An effective binaural processing model based on interaural phase differences

Von der Fakultät für Mathematik und Naturwissenschaften
der Carl-von-Ossietzky-Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation

Dipl.-Phys. Mathias Dietz
geboren am 19. Juni 1979
in Bad Wildungen

Gutachter: PD Dr. Volker Hohmann

Zweitgutachter: Prof. Dr. Dr. Birger Kollmeier

Tag der Disputation: 28. Oktober 2009

Abstract

Humans have the ability to blindly localize sound sources (Sanchez-Longo et al., 1957) and can focus on one out of several concurrent talkers in order to understand exactly the desired talker (Cherry, 1953). One important reason for these phenomena is the ability of the human auditory system to estimate interaural differences by comparing signals from the left and the right ear. Based on the interaural differences the directions of sound sources can be estimated. Differences exist in either timing (interaural time difference, ITD or interaural phase difference IPD) or in level (interaural level difference, ILD). The standard model for ITD detection from Jeffress (1948) is based on an internal cross-correlator with two counterdirective chains of delay elements (delay lines). Psychoacoustic data from various recent studies cannot be explained by a delay line model and also physiologic data from the superior olivary complex and the inferior colliculus of guinea pigs does not support the idea of delay lines in mammals (see McAlpine und Grothe, 2003).

The aim of this thesis is to create a binaural processing model with an effective model structure but without delay lines. An effective model structure is a chain of signal processing elements, with the philosophy not to model a specific neural realization but rather to quantify the transmission information. While physiologic models describe for instance a cohort of neural responses as stochastically generated time functions of voltage differences, this may be simplified by an effective model to just one value proportional to the mean activity over all neurons and over time. The new effective model is introduced in chapter 2. It derives the interaural timing disparities by instantaneous phase comparison instead of delay lines and is therefore referred to as the “IPD model”.

Chapter 3 describes a comparison between the new IPD model and an excitatory-inhibitory (EI) model based on subtraction between counterdirective delay lines (Breebaart et al., 2001a). The focus of the comparison is on temporal resolution. For this a psychoacoustic study on broadband binaural beat detection was conducted. Beating could be detected even above 200 Hz beat frequency. This high limit of temporal resolution could be modeled best without any integration time constant because the auditory filter bandwidth already accounts for the limitation.

In the fourth chapter the IPD model is extended to a lateralization model. This model was evaluated with data from a psychoacoustic study on the lateralization of sinusoidally amplitude modulated (SAM) tones. With consideration of the neural activity as a function of IPD (McAlpine et al., 2001; Marquardt and McAlpine, 2007) the data could be modeled. The model was also able to account for previously published data (Trahiotis and Stern, 1989). In a cross correlation model the same data led to contradictions with either physiology or psychoacoustics depending on the length of the delay lines (Thompson et al., 2006).

Furthermore the IPD model was tested for direction of arrival estimation with several concurrent speakers (chapter 5). Up to five stationary concurrent speakers could be localized in free field simultaneously with errors < 5°. The model structure and the limitation to interaural features humans can extract lead to feature extraction algorithms different to those of technical approaches. One important example is the coherence filter that is facilitated by a higher temporal resolution.

In summary the IPD model has proven as a successful approach, to overcome the restrictions of cross-correlation based models for a broad range of issues.

Zusammenfassung

Menschen verfügen über die Fähigkeit, Schallquellen blind zu orten (Sanchez-Longo u. a., 1957) und sich in einem Stimmengewirr spezifisch auf einen Sprecher zu konzentrieren und diesen dadurch zu verstehen (Cherry, 1953). Ein wichtiger Grund für diese Phänomene ist die Fähigkeit des Hörsystems, die Signale von linkem und rechtem Ohr zu vergleichen. Aus diesen interauralen Unterschieden kann die Richtung von Schallquellen geschätzt werden. Interaurale Unterschiede sind zum einen Zeit- bzw. Phasendifferenzen (ITD, engl. interaural time difference bzw. IPD, engl. interaural phase difference) und zum anderen interaurale Pegeldifferenzen (ILD, engl. interaural level difference). Im Standardmodell zur ITD-Bestimmung von Jeffress (1948) wird von einem internen Kreuzkorrelator mit zwei gegenläufigen Ketten von Verzögerungselementen ausgegangen. Psychoakustische Daten von verschiedenen neueren Studien lassen sich jedoch nicht durch dieses Modell erklären und auch neurophysiologische Daten aus dem superioren Olivenkomplex und dem Colliculus Inferior von Meerschweinchen sprechen nicht für Verzögerungsketten in Säugetieren (siehe z.B. McAlpine und Grothe, 2003).

Ziel dieser Arbeit ist es ein effektives binaurales Modell zu erstellen, ohne die Verwendung von Verzögerungsketten. Unter einem effektiven Modell ist eine Abfolge aus elementaren Signalverarbeitungsschritten zu verstehen, wobei das Signal keine konkrete Realisierung beschreibt, sondern die übertragene Information quantifiziert. Während beispielsweise ein physiologisches Modell eine Schar neuronaler Antworten als stochastische Zeitfunktionen von Potentialdifferenzen nachbildet, vereinfacht ein effektives Modell dies etwa durch einen einzigen Wert, der proportional zum Schar- und Zeitmittel ist. In Kapitel 2 wird das neue effektive Modell eingeführt. Es bestimmt zeitliche Unterschiede auf Basis von instantanem Phasenvergleich und ohne Verzögerungsketten und wird im Folgenden als „IPD Modell“ bezeichnet.

Kapitel 3 behandelt einen Vergleich zwischen dem neuen IPD Modell und einem exzitatorisch-inhibitorischen (EI) Modell bezüglich der Zeitauflösung. Das EI Modell basiert auf Differenzvergleich zwischen den gegenläufigen Verzögerungsketten (Breebaart u. a., 2001a). Dafür wurden psychoakustischen Messungen zur Detektion von breitbandigen binauralen Schwebungen durchgeführt. Die binauralen Schwebungen konnten bis über 200 Hz detektiert werden. Diese hohe Grenze der binauralen

Zeitauflösung konnte am besten ohne zeitliche Integrationskonstante modelliert werden, da die Beschränkung durch die Bandbreite der auditorischen Filter bereits die Daten erklärt.

In Kapitel 4 wurde das IPD Modell zu einem Lateralisationsmodell erweitert. Dieses erweiterte Modell wurde an parallel durchgeführten psychoakustischen Messungen zur Lateralisation von sinusförmig amplitudenmodulierten (SAM) Tönen evaluiert. Durch die Berücksichtigung der neuronalen Aktivität als Funktion der IPD (McAlpine u. a., 2001; Marquardt und McAlpine, 2007) konnten die Daten modelliert werden. Das Modell eignet sich ebenso zur Erklärung von Literaturdaten (Trahiotis und Stern, 1989), die im Kreuzkorrelationsmodell je nach Länge der Verzögerungsketten entweder im Widerspruch zur Physiologie oder zur Psychoakustik stehen (Thompson u. a., 2006).

Außerdem wurde das IPD Modell auf seine Eignung zur Richtungsbestimmung mehrerer gleichzeitiger Sprecher untersucht (Kapitel 5). Bis zu fünf stationäre Sprecher konnten unter idealen Bedingungen gleichzeitig mit einem Fehler $< 5°$ bestimmt werden. Die Struktur des Modells und die Einschränkung auf die von Menschen extrahierbaren interauralen Informationen erforderten zum Teil andere Verfahren zur Merkmalsextraktion als technische Ansätze. Ein wichtiges Beispiel hierfür ist der Kohärenzfilter, dessen Leistung auf der hohen zeitlichen Auflösung beruht.

Zusammenfassend hat sich das IPD Modell als erfolgreicher Ansatz erwiesen, um für verschiedenste Problemstellungen die Beschränkungen der Kreuzkorrelationsmodelle aufzuheben.

Contents

1. General Introduction 11

2. The IPD model 15

2.1 Monaural pre-processing 15

2.2 Extraction of interaural phase differences 17

2.3 Extraction of interaural level differences 21

3. Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences 23

3.1 Introduction 24

3.2 Binaural processing models 26

3.2.1 Stages of monaural pre-processing 27

3.2.2 Binaural processing based on IPD rate coding 28

3.2.3 Binaural processing based on EI-cells 32

3.3. Psychoacoustic measurements 33

3.3.1 Method 33

3.3.2 Results 35

3.4. Simulations 37

3.4.1 Phase-warp 37

3.4.2 Noise and pure tones with ITD 40

3.4.3 N_0S_π -detection 41

3.4.4 Transposed stimuli 42

3.5 Discussion 43

3.6 Conclusions 46

4. Lateralization of stimuli with independent fine-structure and envelope based temporal disparities 49

4.1 Introduction 50

4.2 Model Structure 53

4.3 Methods 59

4.3.1 Subjects ... 59
4.3.2 Apparatus and stimuli ... 59
3.3 Procedure ... 61
4.4 Experimental results and model predictions ... 63
4.4.1 Experiment 1: Matching of carrier IPD ... 63
4.4.2 Experiment 2: Matching of modulator IPD ... 65
4.4.3 Experiment 3: Trading of carrier and modulator IPD ... 65
4.4.4 Experiment 4: Trading of carrier and modulator IPD with a two-tone complex ... 67
4.4.5 Experiment 5: Level dependence of lateralization ... 68
4.4.6 Accounting for published data ... 70
4.5 Discussion ... 70
4.5.1 Psychoacoustic results ... 70
4.5.2 Modeling lateralization ... 72
4.5.3 Physiological validity of the model ... 74
4.5.4 Relation of experimental to modeled data ... 76
4.6 Conclusions ... 78

5. Model-based direction estimation of concurrent speakers from a binaural signal ... 81
5.1 Introduction ... 82
5.2 Model structure ... 86
5.2.1 Auditory processing ... 86
5.2.2 Feature selection ... 89
5.3 Method ... 92
5.3.1 Signal generation ... 92
5.3.2 IPD-to-azimuth-direction mapping function ... 93
5.3.3 Across filter integration ... 93
5.3.4 Task-dependent direction estimation ... 94
5.4 Model evaluations ... 95
5.4.1 Temporal resolution ... 95
5.4.2 Influence of modulation IPDs ... 97
5.4.3 Influence of noise and reverberation ... 98
5.4.4 Tracking of moving speakers ... 100

5.5 Discussion....102
5.6 Conclusions....106

6. Summary and concluding remarks....107

Bibliography....111

Chapter 1

General Introduction

Animals and humans benefit from the ability to determine the direction of a sound source. Owls can localize mice rustling with leaves in darkness thanks to their exceptional sound localization. Humans can hear from which direction a car or a potential danger is approaching. Especially blind people often develop the ability to extract a lot of information about their surrounding space by spatial hearing.

It is long known that the differences in arrival time and in intensity between the left and the right ear are the dominating cues for direction estimation (Thompson, 1882; Rayleigh, 1907). These binaural or interaural differences contain information about the azimuth (α) of the sound source. The travel time of sound from the source to each of the two ears is different (except for sources which originate from straight ahead, $\alpha = 0°$, or behind, $\alpha = 180°$), resulting in an interaural time difference (ITD). The highest ITDs occur for stimuli from the right hand side ($\alpha \sim +90°$) or the left hand side ($\alpha \sim -90°$). The resulting so-called physiological range is about 700 µs for humans and only about 100 µs for guinea pigs with much smaller heads. In addition to the difference in travel time the two received signals may have an interaural level difference (ILD) due to a damping of the averted ear input by the head.

Thompson (1877) reasoned that the binaural sensitivity he observed is probably not due to acoustic interference via the Eustachian tubes but rather to neural comparison in the brain. 70 years later Jeffress (1948) came up with the first model of the neural ITD coding. In the Jeffress model the signals from each ear are successively delayed on counterdirective pathways (dual delay lines). Along the dual delay line the two differently delayed signals are compared by coincidence neurons that detect how well the internal delay at each position compensates for the external ITD. It is assumed that the internal delay of each position is known or learned by the auditory system, so that the position of the best synchronized coincidence neuron is a measure for the ITD. This mechanism is closely related to a cross-correlation where the position of the maximum indicates the delay between the two input signals. Since synchronicity is found when the internal delay compensates for the ITD, this principle is also referred to as compensatory delay lines. With the input signal split into frequency channels at the level of the inner ear, the Jeffress model assumes dual delay lines for each frequency channel, all covering at least the physiological range.

The elegance and the analogy to standard cross-correlation procedures helped for the big success and continued acceptance of the Jeffress model for more than 60 years. At the time the model was developed no neurophysiologic knowledge about the auditory system was available and the idea of delay lines with coincidence detection could be neither proved nor disproved by physiology. At about the same time when Jeffress published his model, binaural psychoacoustic studies emerged (e.g., Hirsh, 1948; Blodgett et al., 1956; Pollack and Trittipoe, 1959; Sayers, 1964). The concept of cross-correlation was very successful to explain these data and gave rise to extensions and quantifications of the Jeffress model (e.g., Sayers and Cherry, 1957).

Colburn (1977) extended the Jeffress model by considering monaural neurophysiologic data from the auditory nerve of cats (Kiang, 1968). Colburns highly successful extension also demonstrated that it is important and helpful to include physiologic data from animals in models for the human auditory system. But up to this time there was still no physiologic data addressing the existence of delay lines. This changed with the work by Carr and Konishi (1988, 1990) who found indeed axonal delay lines in the brain stem of barn owls proving the hypothesis of Jeffress at least for this species. It was known that the auditory system of birds and especially of owls differs significantly from the auditory system of mammals and that studies on cats revealed a more heterogeneous response pattern which could not be directly interpreted as delay lines (e.g., Batra et al., 1997), but the Jeffress model remained the de-facto standard for all species (see e.g., McAlpine and Grothe, 2003 for a review). However, several recent experiments in physiology (McAlpine et al., 2001), in functional magnetic resonance imaging (Thompson et al., 2006), in evoked potential measurements (Riedel and Kollmeier, 2006) and in psychoacoustics (Phillips, 2008; Furukawa, 2008) indicate that there are difficulties to interpret the data with a Jeffress model.

Without going into detail it is necessary to review some physiologic data of ITD sensitivity in the mammalian brain stem. There are two primal nuclei that receive input from both ears, namely the medial superior olive (MSO) and the lateral superior olive (LSO). While these two nuclei were traditionally thought to be specialized for one interaural parameter (MSO for ITD and LSO for ILD), Joris and Yin (1995) also found ITD dependent responses in LSO neurons, in particular a sensitivity to ITDs in the envelope of high-frequency (> 2 kHz) stimuli. In contrast, low-frequency response functions measured in the MSO (Brand et al., 2002) and in the inferior colliculus (McAlpine et al., 2001) indicate a dominant dependence on the interaural phase difference (IPD) of the fine-structure of low-frequency (< 1.5 kHz) stimuli. The difference between IPD and ITD is important if different frequency channels are considered: The delay range in an ITD detector covers the physiological range, which is almost independent on frequency. An IPD detector does not depend on the physiological range and can detect delays up to the half of the cycle duration of the respective frequency. This translates into a maximum ITD of only 500 µs at 1 kHz but to 5 ms at 100 Hz. Especially the short maximum at 1 kHz, which is already below the physiological range of humans, is in conflict with the Jeffress model.

With the physiologic indication for a separate processing of temporal fine-structure in the MSO and of ILD and temporal envelope processing in the LSO it is also possible to account for the studies mentioned in the penultimate paragraph. All of these studies ask implicitly or explicitly for a new binaural model, either because they demonstrate that a single temporal channel is not sufficient or because they are in conflict with the compensatory delay arrays.

Beside the conflicts of the mentioned studies, lots of other recent psychoacoustic studies (e.g., Beutelmann and Brand, 2006; Culling, 2007) but also modeling studies (Faller and Merimaa, 2004; Calmes et al., 2007) still favor approaches based on Jeffess-like delay compensation. As long as the model is not concerned with physiology there are good reasons to apply delay compensation models because there are the smart extensions and variations of these models (e.g., Lindemann 1986; Stern et al., 1988; Stern and Shear, 1996; Breebaart et al., 2001a; Liu et al., 2001) that have evolved together with many psychoacoustic studies (e.g., Trahiotis and Stern, 1989). These extended models are usually very suitable to account for the data. An additional reason favoring the existing delay compensation models over more physiologic models is that the signal processing is more comprehensible, sometimes even analytically (e.g., Durlach, 1963), whereas the physiologic processing scheme is still a subject of debate and of fundamental research.

There are also existing binaural models based on interaural phase differences (e.g., Kollmeier and Koch, 1994; Borisyuk et al., 2002; Nix and Hohmann, 2006). These models are rather focusing on a specific issue such as modulation IPD perception for speech enhancement (Kollmeier and Koch, 1994) or sound source localization based on empirical statistics (Nix and Hohmann, 2006). The IPD based models were therefore not employed to model standard binaural psychoacoustic data such as binaural masking release or complex tone lateralization.

The aim of this thesis is to develop a binaural processing model which is not in conflict with mammalian physiology but also sufficiently comprehensive, and flexible enough to account for a vast range of psychoacoustic data. Another requirement for the model is that it can be used for applications such as sound source localization.

Chapter 2 introduces the core of the developed model before three applications are presented in the chapters 3-5. Fortunately, the extraction of the interaural functions, which was developed before most of the applications, has proven to be suitable for later experiments. However, most applications require a special feature extraction from the interaural functions, e.g. time dependent vs. averaged over time, frequency selective vs. integrated over frequencies. Therefore the specific feature extraction is different in each application and presented in the respective chapter.

The applications deal with temporal aspects and binaural masking release (chapter 3), lateralization of complex stimuli (chapter 4) and speaker localization in adverse conditions (chapter 5).

Chapter 2

The IPD model

The demand on the model is both comprehensiveness of the processing and no principal conflict with mammalian physiology. Such a comprehensive modeling concept with subsequent processing blocks, each one built of standard elements of digital signal processing is called “effective modeling” (e.g., Dau et al., 1996; Breebaart et al., 2001a, 2001b, 2001c). Each anatomical element of the auditory periphery can be accounted for by one specific building block, whose parameters are generally deducted from direct measurements of the input-output characteristics of the element. The resulting monaural pre-processing model is therefore adapted from Dau et al. (1996) and briefly described in Sec. 2.1. In the central auditory system it is harder to assess the input-output characteristic of each nucleus separately, especially since efferent connections cause active non-linear responses. The effective binaural model developed in Sec. 2.2 therefore describes a hypothetical processing mechanism motivated by the physiologic indications of the potential absence of a place coding map via counterdirective delay lines (Jeffress, 1948), the half-cycle limitation of maximal response from binaural neurons (McAlpine et al., 2001) and the indication for a separate processing of temporal fine-structure and temporal envelope (e.g., Furukawa, 2008).

The aim of this chapter is to introduce a novel binaural interaction model in detail. This model is the basis for the applied simulations of the main chapters 3-5. These chapters all contain a section which is reintroducing a specific version of this model. It cannot be avoided that there is a big overlap, but the current chapter is the most detailed model introduction with a focus on the general aspects.

2.1 Monaural pre-processing

The stages of peripheral pre-processing are briefly described in this section in the processing order. Parameter values are sometimes left out because they may vary in different applications:

1. The middle ear transfer characteristic was modeled using a time-invariant bandpass (e.g., Breebaart et al., 2001a).
2. In order to model the frequency analysis performed by the basilar membrane a linear fourth-order gammatone filterbank (Patterson et al., 1987) was employed with an analytic filter output in the implementation of Hohmann (2002). A transfer function for some low-frequency channels is shown in Fig. 2.1.

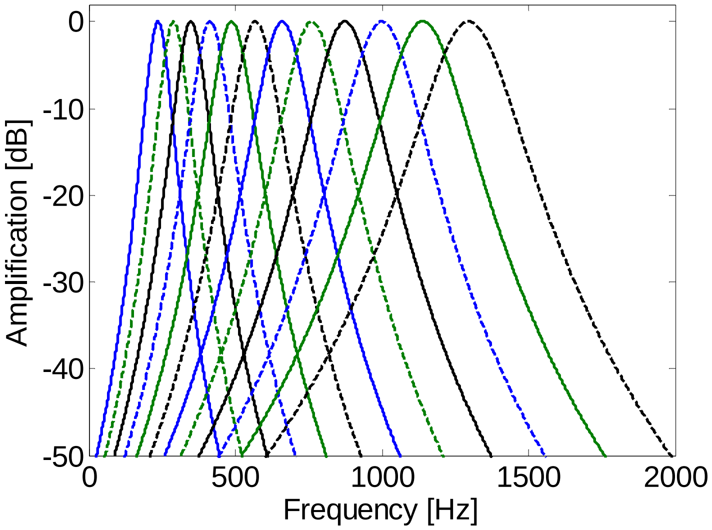

Fig. 2.1 - Transfer functions of 12 low-frequency bands from the gammatone filterbank.

3. Cochlea compression was accounted for by instantaneous compression with a power of 0.4 (e.g., Ewert and Dau 2000, Ruggero and Rich 1991).
4. The mechano-electrical transduction process in the inner hair cells was modeled with a half-wave rectification with a successive 770-Hz fifth-order low-pass filter as used by Breebaart et al. (2001a). The influence of this stage is illustrated in Fig. 2.2. The half-wave rectification leads to the demodulation of the amplitude modulations to low frequencies. The low-pass filter results in the partial suppression of the original 1-kHz peak with respect to the low-frequency demodulation.
5. A finite hearing threshold can optionally be employed by adding uncorrelated Gaussian noise.

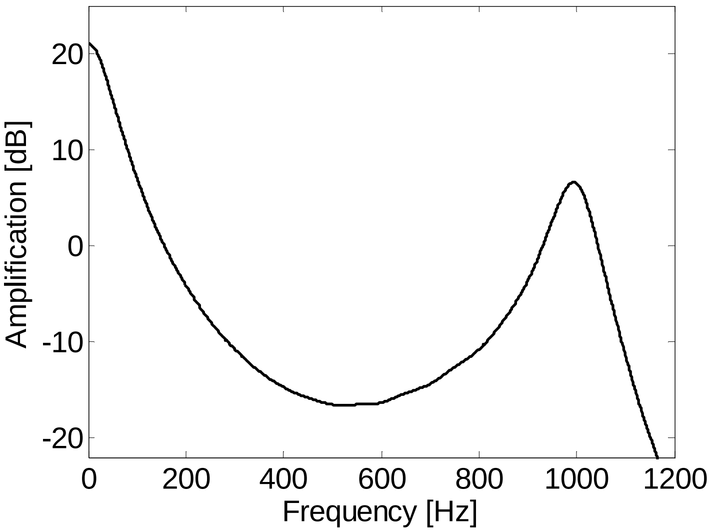

Fig. 2.2 - Transfer function of the monaural processing up to the mechano-electrical transduction. As an example only the 1-kHz band was selected. Two regions of high spectral energy can be identified: (1) demodulated by the half-wave rectification modulation or envelope frequencies appear from 0 up to about 400 Hz and (2) spectral density of fine-structure information peaks at 1 kHz. The positive amplification is caused by the instantaneous compression.

2.2 Extraction of interaural phase differences

The frequency dependent ITD-range limitation of binaural neurons to the half of the cycle duration at the respective characteristic frequency is a strong motivation to extract IPDs instead of ITDs in the first place. In free-field listening, however, ITDs and not IPDs are almost constant over frequency. This difference is not a contradiction. The output of the primary IPD detector can easily be labeled as ITD or as azimuth before further integration across frequencies.

In order to determine an IPD, it is necessary that the outputs $m_{\mathrm{left}}^{cf}(t)$ and $m_{\mathrm{right}}^{cf}(t)$ of the monaural pre-processing at the center frequency cf provide sufficient information about the phases of the signals. However, the waveform and spectral shape of the signals processed by the hair cells are altered by the half-wave rectification. In comparison to the bandlimited signals after the peripheral gammatone filterbank (Fig. 2.1), a hair cell processed band (Fig. 2.2) has a broadened spectrum including a DC component, the demodulated envelope, and usually energy in the frequency region of the original bandlimited signal. As illustrated in the two right panels in Fig. 2.3, the hair cell output cannot be represented by a monotonic phase, which is required for a stable output in the binaural stage. For extracting stable and reasonable IPDs, both left and right signals must always have the same direction of rotation in the complex plane. In practice, this requires an unchanged direction of rotation around the center of the complex plain for each signal which is called a monotonic phase. Changes in the direction of rotation are caused by low frequency components and particularly by adding a DC component, which shifts the center of rotation away from the center of the complex plane.

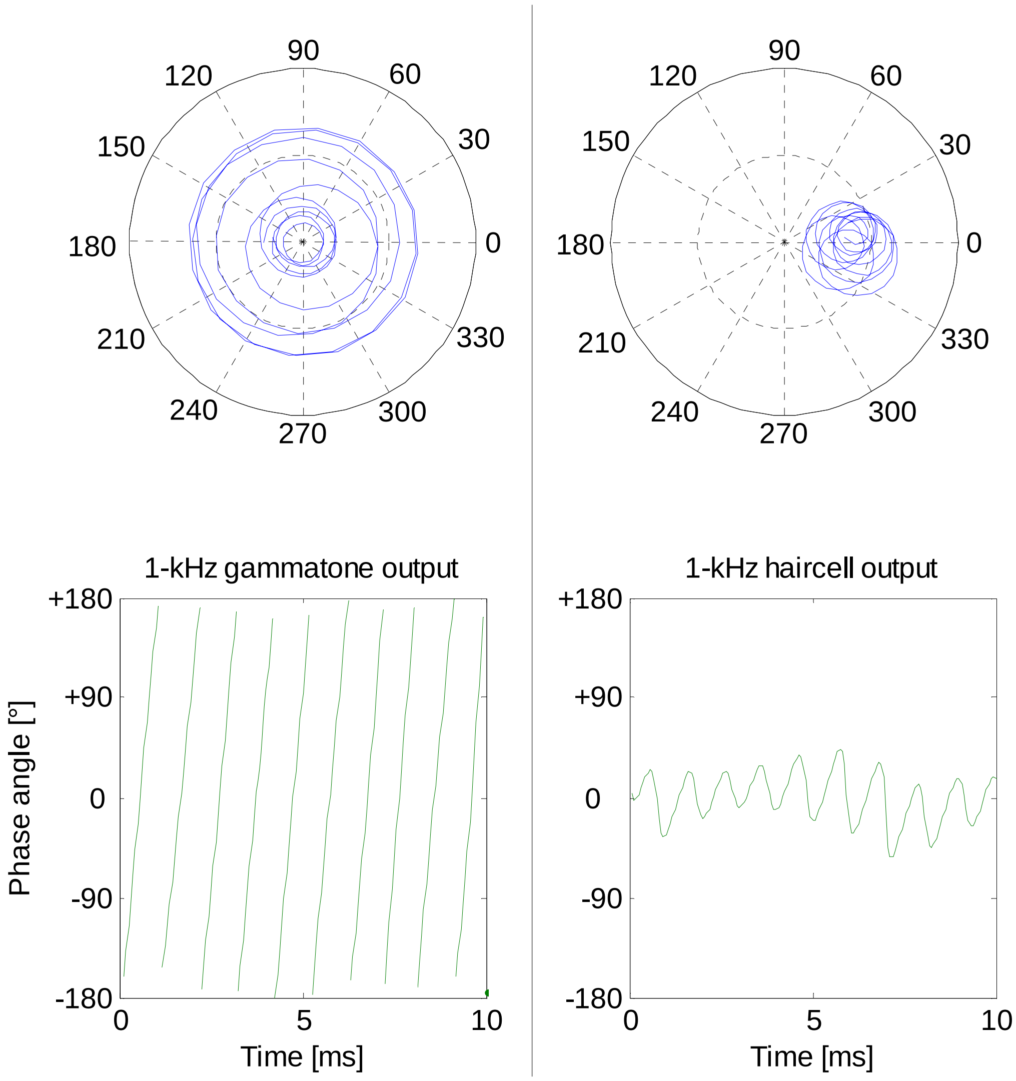

Fig. 2.3 - Exemplary 10-ms speech segment after gammatone processing (left) and the analytical output of the hair cell transduction stage (right) in the 1-kHz band. In the upper panels the analytical signals are plotted in polar coordinates. The lower panels show the isolated phase angle at each stage. The half-wave rectification shifts the signal out of the center of the complex plane which results in the non-monotonic phase.

The computational necessity to further process the hair cell output before determining an IPD and the strong indication for separate processing of temporal fine-structure and temporal envelope differences described in chapter 1 (e.g., Furukawa, 2008), both suggest separating the hair cell output into different spectral regions before binaural interaction. For this, each hair cell output channel, characterized by the center frequency *cf* from the preceding gammatone filterbank, is filtered again by two different band-pass filters. The band-pass filters are realized as complex-valued gammatone filters (see Hohmann, 2002 for implementation details). One filter, centered at *cf*, is effectively extracting the fine-structure of the input stimulus in the respective channel.

The other filter which is used in parallel to the fine-structure filter extracts the envelope of the signal. It is called "modulation filter" and is usually set to the dominant modulation frequency between 25 and 300 Hz. In principle several modulation filters can be used in order to cover the complete modulation frequency range.

The two different stages of parallel filtering, i.e. the peripheral filtering at the level of the basilar membrane and the more central separation of each band in two sub-bands for fine-structure and modulation filters, are labeled as follows:

The sub-bands originating from the left ear are labeled $g_{\mathrm{left}}^{cf,type}(t)$ where *cf* indicates the center frequency from the basilar membrane filtering and *type* indicates the second filter (either "fine" for fine-structure or "mod" for modulation). Sub-bands originating from the right ear are labeled $g_{\mathrm{right}}^{cf,type}(t)$ accordingly.

The output of each filterband is fed into the interaural processor together with its counterpart from the other ear. The binaural processing is described in the following for one pair of processed filterbands: $g_{\mathrm{left}}^{cf,type}(t)$ and $g_{\mathrm{right}}^{cf,type}(t)$. The following general description is the same for all bands and types, so that the indices *cf* and *type* do not play a role and will be omitted. Left and right channels are labeled with g_l and g_r respectively.

Since the outputs of the gammatone filters are complex valued, the instantaneous phases $\phi_l(t)$ and $\phi_r(t)$ are given explicitly in the polar representation:

$$g_l(t) = a_l(t) \cdot e^{i\phi_l(t)} \; , \tag{2.1}$$

where $a_l(t)$ is the instantaneous amplitude of the signal $g_l(t)$. The right channel is expressed accordingly. Now a simple subtraction of the phases $\phi_l(t)-\phi_r(t)$ is the only binaural processing step, directly resulting in the interaural phase difference.

However, in the actual implementation of the model, the phase difference is determined by the argument of the complex interaural transfer function (ITF):

$$\mathrm{ITF}(t)=g_l(t)\cdot\overline{g_r}(t)=a_l(t)\cdot a_r(t)\cdot e^{i(\phi_l(t)-\phi_r(t))}\,, \tag{2.2}$$

where $\overline{g_r}(t)$ is the complex conjugate of $g_r(t)$.

The advantage of using the ITF is the possibility to apply an averaging low-pass filter, with each phase difference value weighted with the product of left and right amplitude. The low-pass filtering at this stage is used by binaural models to account for a finite temporal resolution. The IPD can then be extracted from the low-pass filtered ITF by

$$\mathrm{IPD}(t)=\arg\left([\mathrm{ITF}(t)]_{lp}\right), \tag{2.3}$$

with “*lp*” denoting the optional low-pass filter.

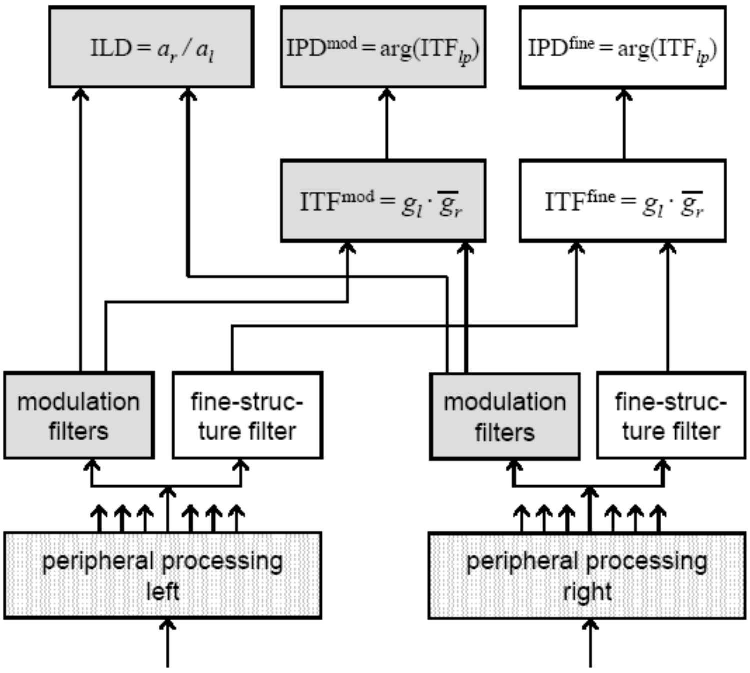
ILD = a_r / a_l
IPD$^{\mathrm{mod}}$ = arg(ITF$_{lp}$)
IPD$^{\mathrm{fine}}$ = arg(ITF$_{lp}$)
ITF$^{\mathrm{mod}}$ = $g_l \cdot \overline{g_r}$
ITF$^{\mathrm{fine}}$ = $g_l \cdot \overline{g_r}$
modulation filters
fine-structure filter
modulation filters
fine-structure filter
peripheral processing left
peripheral processing right

Fig. 2.4 - Processing stages of the IPD model. White boxes indicate fine-structure processing which is usually found in the medial superior olive, dark grey boxes indicate processing of modulation frequencies in the lateral superior olive.

2.3 Extraction of interaural level differences

The modulation band-pass filters still do not cover the very lowest modulation frequencies and static differences between a pair of left and right signal channels. However, these differences are exploited by the human auditory system in the form of interaural level differences (Halverson, 1922). In order to include these cues in the framework of the IPD model, the modulation filters need to be amended by low-pass filters. These filters operate in parallel to the fine-structure and the modulation band-pass filters and represent a third version called “low” for the filterband argument *type*. The ILD can now be derived from the energy ratio of a pair of modulation low-pass filter outputs $g_{\text{left}}^{cf,\text{low}}(t)$ and $g_{\text{right}}^{cf,\text{low}}(t)$:

$$\text{ILD}^{cf}(t) = \frac{20}{c} \cdot \log_{10}\left(\frac{\left| g_{\text{right}}^{cf,\text{low}}(t) \right|}{\left| g_{\text{left}}^{cf,\text{low}}(t) \right|} \right). \tag{2.4}$$

The ILD is expressed in dB and therefore divided by the compression exponent c in order to scale the internal representation to the original ILD between the ears prior to basilar membrane compression.

A sketch of the IPD model is shown in Fig. 2.4. The reason why low-pass and band-pass modulation filters are fused to one block “modulation filters” is based on physiologic evidence that neurons in the lateral superior olive are sensitive to ILD and temporal envelope differences (e.g., Joris and Yin, 1995).

Previously published delay lines models were forced by design to separate cues based on timing from cues based on level differences (e.g., by two orthogonal dimensions: Breebaart et al., 2001a). Beside the absence of delay lines, it is probably the most important difference of the current approach to separate by spectral content of the demodulated hair cell output and to assign temporal differences in the stimulus envelope as a separate cue, processed independently from fine-structure IPD.

Chapter 3

Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences[1]

Abstract

A model of the effective processing of interaural timing disparities in the human auditory system is presented which provides modifications and extensions to existing models motivated by recent physiological findings. In particular, an established model of excitatory-inhibitory (EI) neuronal connectivity is complemented by a model that is based on a rate code derived from the interaural phase difference (IPD). The IPD model is shown to successfully simulate literature data on fine-structure and envelope-based binaural detection and lateralization experiments. In order to investigate the processing of temporal fluctuations of interaural timing disparities, detection thresholds of broadband binaural-beat stimuli were measured in six normal-hearing listeners and were compared with model simulations. In a first experiment, the highest detectable beat frequency was found to be 96 Hz for a noise bandwidth of 550 Hz and 219 Hz for a bandwidth of 1100 Hz. Both models predicted lower thresholds, but performed increasingly better when the integration time constants of the binaural processors were reduced. In a second experiment, the signal-to-noise ratio at the detection threshold of binaural-beat stimuli mixed with interaurally uncorrelated noise was measured as a function of the beat frequency. The threshold increased about 1.7 dB per octave which was simulated similarly by both models. The results indicate that the primary temporal resolution of the binaural system for detecting interaural timing disparities is much higher than the temporal resolution found in higher auditory processes as supposedly involved in, e.g., masking.

---

[1] Reprint of Brain Research (vol. 1220, p. 234-245, 2008). Same title. Copyright Elsevier.
Mathias Dietz, Stephan D. Ewert, Volker Hohmann and Birger Kollmeier.
With permission by Elsevier.

3.1 Introduction

In binaural sound perception with headphones, the intracranial locus of a sound source is dependent on both interaural timing- and interaural level differences. In natural situations, such differences arise from different travel times of a sound from an external sound source to each ear and filtering effects of the head, upper torso and pinnae, resulting in interaural timing and level differences (e.g., Mehrgardt and Mellert, 1977). The combination of both types of interaural differences mediates spatial hearing. It is commonly assumed that in humans interaural timing differences are important at low frequencies (< 1.5 kHz) whereas interaural level differences become increasingly important at high frequencies where the influence of the head shadow is more prevalent (Duplex theory of binaural hearing, Rayleigh, 1907).

The dominance of interaural timing differences at low frequencies is presumably related to the high degree of phase-locking found in auditory nerve (AN) responses. The ability of AN fibers to temporally synchronize to the stimulus fine-structure usually diminishes at a few kHz, depending on the animal model (Palmer and Russell, 1986). For high frequencies, static interaural differences in the stimulus envelope (interaural level differences) are preserved as well as the temporal structure of the envelope. Psychophysical data provide evidence that interaural timing disparities carried in the envelope of high-frequency carriers can be detected and do provide timing cues for lateralization (Van de Par and Kohlrausch, 1997). A popular neuronal model of the processes underlying the perception of interaural timing differences is the so-called "Jeffress model" (Jeffress, 1948). The model consists of several coincidence detectors for the signal paths from both ears, which receive their input signals along two opposed internal delay lines. In such a configuration, the external interaural time difference (ITD) is determined by the position along the delay line at which the internal delays just compensate for the external delay and the highest degree of coincidence occurs. With this process, a neuronal place mapping (position of the coincidence neuron) of interaural timing differences is realized. It has been shown that the neuronal concept of coincidence detection (excitatory-excitatory, EE) combined with a delay line is conceptually similar to the mathematical operation of cross-correlation (e.g., Batra and Yin, 2004). Many studies of binaural processing are therefore based on interaural cross-correlation as synonym of the Jeffress model (e.g., Lindemann 1986) or motivated by its topology. A modern variation of the neuronal Jeffress model is based on contralateral inhibition (Breebaart et al., 2001a). This model is based on an excitatory-inhibitory (EI) neuronal connectivity between both ears and is therefore termed EI model in the following. Instead of short-term correlation for coincidence detection (EE; Colburn and Durlach, 1978), the EI model subtracts the differently delayed stimuli and uses the reduction of neural firing as a cue.

Although the Jeffress model is appealing in its function, there is still an ongoing discussion about its existence in different species (Harper and McAlpine, 2004). Physiological findings in the barn-owl have revealed that this species uses a "Jeffress-like" delay line (Carr and Konishi, 1990). However, physiological investigations of guinea pigs do not unambiguously support the implementation of delay lines in mammals (McAlpine et al., 2001).

An alternative concept to the delay line might be a neuronal population- or rate coding of interaural time differences (Fitzpatrick et al., 1997, McAlpine and Grothe, 2003), where the firing rate of a neuron is generally an injective function of the ITD. Hence, the ITD can unambiguously be determined by a given firing rate. Brand et al. (2002) have shown in the gerbil that at low frequencies, firing rates increased even for time delays much longer than the highest possible ITD resulting from a free field sound source for a given species. This leads to an increasing firing rate of a neuron in the limits of the physiological range, roughly determined by the distance between the ears of the species, and a peaking of the firing rate function outside of this range. Although the exact peak of the function was found to be dependent on the best frequency of the neuron (Brand et al. 2002), the interaural phase difference (IPD) at which peak firing occurs is rather constant. Most neurons showed a steep increase in their firing rates for small ITDs or IPDs and they all reached their highest rate for phase shifts not higher than $\pi/2$. These investigations reveal physiological evidence for rate coding in the medial superior olive (MSO).

While the findings of the above studies suggest an alternative strategy of the coding of interaural timing, it might be possible that the suggested rate coding coexists with delay lines or other coding strategies in order to exploit interaural timing disparities (Harper and McAlpine, 2004). When assuming a rate-code that peaks at a fixed IPD independently of the best frequency, however, the firing rate can be universally expressed by the IPD.

Taken together, there are physiological findings supporting the fundamental role of interaural phase differences in binaural processing of the auditory system (e.g., Marquardt and McAlpine, 2007). Based on these findings, the first objective of this paper is to examine whether the concept of IPD-based rate coding can be embedded in an existing framework of effective modeling (e.g., Dau et al., 1996; Breebaart et al., 2001a). Predictions of the IPD model are compared to predictions of the EI model by Breebaart et al. (2001a).

Furthermore, a recent physiological and psychoacoustic study by Siveke et al. (2007) has revealed that the binaural system is sensitive to fast binaural fluctuations up to 256 Hz. In their study, they used a broadband binaural-beat stimulus, called phase-warp, to estimate the temporal resolution of the binaural system. The broadband stimulus can be thought of as a sum of individual tones with a frequency offset between both ears, each producing binaural beats at the rate of the offset frequency. Their findings may have influence on effective binaural modeling since the temporal resolution of established models (e.g., Breebaart et al., 2001a, 2001b, and 2001c) has only been tested against much slower fluctuations. Therefore, the second objective of the paper is to compare the behavior of the IPD model and a processor based on the EI model by Breebaart et al. (2001a) with psychoacoustic measurements of the phase-warp stimulus suggested by Siveke et al. (2007). The third objective of the paper is to demonstrate the basic properties of the IPD model by simulating literature data on lateralization of tones (Sayers , 1964) and noises (Bernstein and Trahiotis, 1982), on the “classical” binaural masking level difference (Hirsh, 1948), and on the transposed BMLD (van de Par and Kohlrausch, 1997).

In the next section, the binaural processing models are introduced. The third section describes psychoacoustic measurements with the phase-warp. The numerical simulation of the measurements is described in Sec. 3.4, together with simulations of literature data. Results are discussed in Sec. 3.5.

3.2 Binaural processing models

In order to provide a fair comparison of different model concepts for binaural processing, the same monaural, peripheral pre-processing is used. The principle concept of effective modeling was taken from Dau et al. (1996), and most specific parameters in the processing stages were taken from the model in Breebaart et al. (2001a). The specific stages are briefly described in the following. Detailed information on the binaural processing model based on IPD rate coding, as suggested in this study, is given along with a short review of the relevant parts of the EI model. An overview of all processing stages is shown in Fig. 3.1.

3.2.1 Stages of monaural pre-processing

In the following list, the stages of peripheral pre-processing are briefly described in the same order as the input signals for both ears are processed:

1. The combined outer- and middle ear transfer characteristic was modeled using a 1-4 kHz band-pass with 6 dB/oct slopes. For the simulation of signals presented over headphones further directional filtering using head-related impulse responses is not required.
2. The frequency analysis performed by the basilar membrane was modeled with a linear fourth-order gammatone filterbank (Patterson et al., 1987; Hohmann, 2002). In addition, an instantaneous compression with a power of 0.4 is used (e.g., Ewert and Dau, 2000). This compression power is consistent with physiological estimates of basilar membrane compression for mid-range levels (Ruggero et al., 1997). For computational simplicity, the compression was implemented after the half-wave rectification in the next stage.
3. To model the mechano-electrical transduction process in the inner hair cells, a half-wave rectification with a successive 770-Hz fifth-order low-pass filter was employed as used in Breebaart et al. (2001a).
4. By adding uncorrelated Gaussian noise to all filterbands after the hair cell transformation, a finite hearing threshold was employed. The noise has the same rms-value as a 0-dB 2-kHz pure tone after half-wave rectification and prior to low-pass filtering. The 2-kHz reference was used since it has the lowest absolute threshold with the given outer- and middle ear filter. With the noise added after the low-pass in the hair cell transduction stage, the signal-to-noise ratio (SNR) for frequency components above 770 Hz is directly influenced.

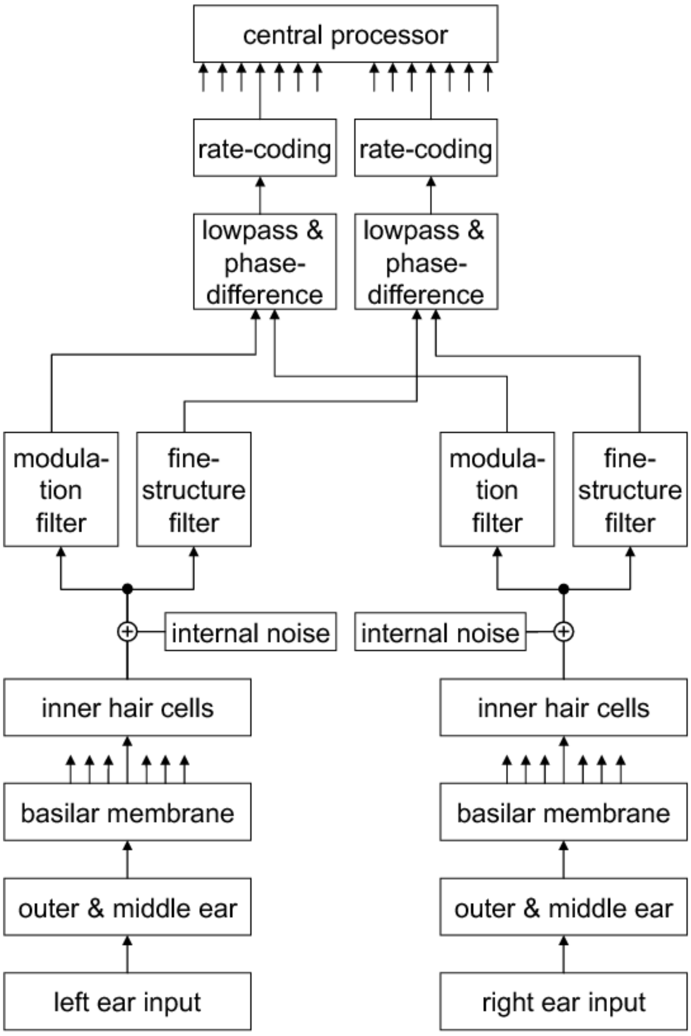

Fig. 3.1 - Successive stages of the binaural model. The multiple arrows behind the basilar membrane simulation symbolize that the signal is split in several frequency bands. The monaural pathways after the hair-cell transduction stage and the addition of internal noise are left out in this sketch. Copyright Elsevier.

3.2.2 Binaural processing based on IPD rate coding

In order to determine a phase difference in the binaural processor, it is necessary that the outputs of the monaural pre-processing at the center frequency *cf*, $m_{\mathrm{left}}^{cf}(t)$ and $m_{\mathrm{right}}^{cf}(t)$ provide sufficient information about the phases of the signals. However, the waveform and spectral shape of the signal processed by the hair cells is altered by the half-wave rectification. In comparison to the bandlimited signals after the peripheral gammatone filterbank, a hair cell processed band has a broadened spectrum including a DC component, the demodulated envelope, and usually energy in the frequency region of the original bandlimited signal. These signals cannot be represented by a monotonic phase, which is required for a stable output in the binaural stage[1]. Hence, an additional spectral limitation of the hair cell stage output is required which is obtained by a second stage of band-pass filtering. In order to extract the whole information in each (peripheral) channel, the hair cell stage output would again have to be processed by a complete filterbank. This processing would result in a two-dimensional array of filterbands with the dimensions (i) peripheral channel and (ii) “modulation filter” channel. Thus, the second filterbank motivated by the requirements of the binaural processing suggested here would be conceptually comparable to a modulation filterbank as suggested earlier by Dau et al. (1997) and Ewert and Dau (2000), ranging up to frequencies similar to the peripheral center frequency of the channel itself. For the model simulations in the present study, however, it was sufficient to process each peripheral channel by only two different band-pass filters. The band-pass filters were realized as complex-valued gammatone filters (see Hohmann, 2002 for implementation details). One filter was centered at the same frequency as the initial, peripheral gammatone filter of the respective channel. Therefore, this filter extracts the remaining fine structure of the input stimulus in the respective channel, as available after the half-wave rectification and low-pass filtering in the model’s hair cell stage. Although the filter could be considered as a “modulation” filter in the notation of former model approaches (Dau et al., 1997), it is referred to as “fine-structure filter” in the following, describing its primary features in the binaural processing context. The order of the filter was set to two and its bandwidth to the half of the respective center frequency (Q factor $Q = 2$).

The other filter which was used in parallel to the fine-structure filter extracts the envelope of the signal. It is called “modulation filter” in agreement with the notation in Dau et al. (1997). However, the phase information (of the envelope) that is required here in order to extract the IPD was completely discarded in Dau et al. (1997) and Ewert and Dau (2000), where the outputs of modulation filters were analyzed for monaural or diotic signal conditions. The modulation filter was centered at 150 Hz for all channels. Again, a complex-valued second-order gammatone filter with $Q = 2$ was employed. The modulation filter extracts a large range of modulation frequencies in the respective peripheral channel, which were demodulated from the envelope of the input stimulus by

---

[1] Usually, the requirement for phase analysis is only a differentiable phase which is practically always given because of the finite upper frequency limit. However, in order to extract useful information from a phase difference, both signals must have (mostly) the same direction of rotation in the complex plane. In practice, this requires an unchanged direction of rotation for each signal which is called a monotonic phase. Changes in the direction of rotation are caused by low frequency components or particularly by adding a DC component which shift the center of rotation away from the center of the complex plane.

the highly nonlinear half-wave rectification in the hair cell stage. $Q = 2$ is the lowest possible Q factor which provided a (mostly) monotonic phase in the tested conditions. The bandwidth of the filters roughly coincides with psychoacoustic estimates of modulation filter bandwidth in Dau et al. (1997) and Ewert and Dau (2000). The center frequency of the modulation filter was set to 150 Hz in order to cover the frequency region in which most experiments have been conducted, and in which the sensitivity to monaural modulation detection is generally still quite high (e.g., Kohlrausch et al., 2000).

Each filterband is unique in the combination of center frequency (*cf*) from the basilar membrane filtering and the type of the second filter (where *type* is either fine-stucture or modulation). The output of each filterband is fed into the binaural processor together with its counterpart from the other ear. The binaural processing is described in the following for one pair of processed filterbands: $g_{\text{left}}^{cf,type}(t)$ and $g_{\text{right}}^{cf,type}(t)$. The parameters *cf* and *type* are constant in the following and will be omitted. Left and right channels are labeled with g_l and g_r respectively.

Since the outputs of the gammatone filters are complex valued, the instantaneous phases $\phi_l(t)$ and $\phi_r(t)$ are given explicitly in the polar representation:

$$g_l(t) = a_l(t) \cdot e^{i\phi_l(t)} \quad , \tag{3.1}$$

where $a_l(t)$ is the instantaneous amplitude of the signal $g_l(t)$. The right channel is expressed accordingly. Now, the simple subtraction of the phases $\phi_l(t) - \phi_r(t)$ is the only binaural processing step, directly resulting in the interaural phase difference.

However, in the actual implementation of the model, the phase difference is determined by the argument of the complex interaural transfer function (ITF), which can easily be derived in two different ways (Nix et al., 2006; Marquardt and McAlpine, 2007):

$$\begin{aligned} \mathrm{ITF}_1(t) &= g_l(t) / g_r(t) = \frac{a_l(t)}{a_r(t)} \cdot e^{i(\phi_l(t) - \phi_r(t))}, \\ \mathrm{ITF}_2(t) &= g_l(t) \cdot \overline{g_r}(t) = a_l(t) \cdot a_r(t) \cdot e^{i(\phi_l(t) - \phi_r(t))} . \end{aligned} \tag{3.2}$$

where $\overline{g_r}(t)$ is the complex conjugate of $g_r(t)$. In order to simulate a finite temporal resolution of the binaural processor, the ITF is low-pass filtered (smoothed). Motivated by the findings of Siveke et al. (2007) showing a relatively high sensitivity to fast interaural timing disparities, a first-order low-pass filter with a cutoff frequency of 64

Hz was used for smoothing. The IPD is then given as the argument of the filtered ITF. An interaural level difference (ILD) can be extracted from the amplitude of ITF_1 only.

The formulae show that the argument of ITF_1 is identical to the argument of ITF_2. However, the low-pass filtering of ITF_1 smoothes out rapid changes in interaural amplitude difference while the smoothing of ITF_2 has a stronger impact on changes in total intensity $a_l \cdot a_r$. Hence there can be minor differences for the smoothed phase difference depending on the way the ITF was derived. All simulations performed in this paper are based on the filtered ITF_2:

$$\mathrm{IPD}(t) = \arg\left(\left[g_l(t) \cdot \overline{g_r}(t)\right]_{lp}\right), \tag{3.3}$$

where $[\ ]_{lp}$ indicates low-pass filtering. However, the choice of the ITF-type has no influence on the results since the differences are small for the signals employed in this study. It is also possible to determine the IPD directly by phase subtraction and to smooth the phase functions before the interaction. In practice, this has almost no influence on the IPD estimate.

The interaural phase difference is the necessary input for the rate-code modeling motivated in the introduction. In a first approximation to data in Brand et al. 2002, the firing rate of neurons in the left MSO is proportional to the sine of the phase difference between 0 and 180° (right signal leading), not considering the noise floor from spontaneous firing. With the neurons in the right MSO reacting symmetrically, the expression

$$l(t) \propto \sin(\mathrm{IPD}(t)) \tag{3.4}$$

does hold for every phase difference, if firing in the right MSO is expressed by negative values (left signal leading). In this notation $l(t)$ denotes the lateralization in the respective band. Left lateralization is expressed by negative values and right lateralization with $l(t) > 0$. For the following analysis, it is assumed that the auditory system can exploit both stationary and time varying features that are preserved in the $l^{cf,type}(t)$- functions.

In order to evaluate the performance of the model, the output function can directly be plotted or analyzed since the model output at a given time is a scalar value. This is a major difference to delay line based models where the primary output of the binaural processor is a vector of neural activity along the delay axis (e.g., Lindemann 1986).

Despite the mathematical elegance of the binaural processor suggested here, it is, of course, questionable whether the auditory system is able to extract the phase or even an analytic signal behind the hair cell processing stage. When assuming bandlimited signals, however, deriving the Hilbert transform is possible using first-order filters (Hohmann, 2002), and a neural circuit for implementing these filters seems possible. Another feasible possibility for a neural circuit to perform a Hilbert transform would be to use two parallel band-pass filters with different phase offsets.

3.2.3 Binaural processing based on EI-cells

For the simulation of an excitatory-inhibitory (EI) processor, the peripheral preprocessing was kept unchanged from the model introduced in the previous section. The binaural processor itself was a special case from the model of Breebaart et al. (2001a). They argued that for the analysis of most conditions it is sufficient to consider only the information for a fixed position on the delay line. The cross-correlation function of the phase-warp stimulus has its peak at $\tau = 0\,\mathrm{ms}$ and thus the delay line output at zero delay displays the highest degree of modulation. Therefore, the EI delay line can be reduced to the squared signal difference, EI_0, which represents the delay line output at zero delay. In order to simulate a finite temporal resolution, Breebaart et al. convolve the output with a double-exponential smoothing window with a time constant of $c = 30\,\mathrm{ms}$, resulting in the following description of the reduced EI processor:

$$\mathrm{EI}_0(t) \propto \left(m_l(t) - m_r(t)\right)^2 * \exp\left(-|t|/c\right) \ , \tag{3.5}$$

where “$*$” denotes convolution. Obviously, the large smoothing time constant results in a severely reduced low-pass frequency (about 3 Hz) compared to the 64-Hz cutoff frequency assumed in the IPD model. In order to compare the models for similar low-pass characteristics, all simulations with the EI model were performed with both the original time constant and a shorter double-exponential smoothing window with a time constant of $c = 1.3\,\mathrm{ms}$, comparable to a 64-Hz low-pass filter.

3.3. Psychoacoustic measurements

Two psychoacoustic measurements were conducted in order to determine the temporal resolution of the binaural system in humans with specific stimulus configurations appropriate for the adjustment of the binaural model suggested in the present study.

3.3.1 Method

Subjects

Seven normal-hearing listeners aged between 24 and 34 years participated in the experiment. All subjects except subject ID had prior experience with other psychoacoustic measurements. Brief training was given to the participants until they were familiar with the procedure and they produced results as stable as subjects MD and SE (authors) who had a few hours of training. Subjects ID, JS, and FZ received a compensation for taking part in the experiment on an hourly basis.

Apparatus and stimuli

The subjects were seated in a double-walled, sound-attenuating booth and listened via Sennheiser HD 580 headphones. Signal generation and presentation during the experiments were computer controlled using the AFC software package for MATLAB, developed at the University of Oldenburg. The stimuli were digitally generated at a sampling-rate of 48 kHz. The transfer function of the headphones was measured with an artificial ear (B&K 4153) and digitally equalized in order to realize a flat amplitude response between 0.1 and 20 kHz.

The stimulus utilized for all experiments was the so-called phase-warp suggested by Siveke et al. (2007). This broadband binaural beat stimulus is generated in the frequency domain using a constant amplitude spectrum and a random phase spectrum for one channel of the stereo stimulus. The second channel is created by shifting the phase spectrum of the whole stimulus cyclically by the beat frequency f_b. This results in a broadband binaural stimulus with each component in the one channel having a respective frequency shifted component in the other channel to produce “classical” binaural beats at a rate of f_b. As in case of the classical binaural beat stimulus, the listener perceives a movement of the sound between both ears for low frequencies (< 10Hz) and a binaural roughness or flutter for higher frequencies. The stimulus is a pure broadband noise when monaurally presented to the either ear, thus providing no monaural modulation or flutter to the auditory system. In order to have a well defined frequency range for the detection of the phase-warp in the experiment, an upper limit f_u was introduced for the phase-warp. In condition 1, the phase-warp was presented alone in the frequency range between 0 Hz and f_u, whereas binaurally uncorrelated noise at the same spectrum level was added in the frequency range between f_u and 24 kHz in condition 2. The 1-s stimuli were presented with 20 ms cosine ramps at a level of 65 dB SPL and inter-stimulus intervals of 0.5 s.

Procedure

A three-interval, three-alternative forced-choice paradigm was used to measure discrimination thresholds between the phase-warp stimulus and binaurally uncorrelated noise with identical white long-term power spectrum. A two-up, one-down procedure was used to estimate the 70.7% correct point of the psychometric function (Levitt, 1971). Subjects had to identify the randomly chosen interval which contained the phase-warp stimulus. In experiment 1a, the adaptive parameter was the beat frequency and an upper cutoff frequency $f_u = 550\,\mathrm{Hz}$ was used. The experiment was repeated for condition 1 and 2, as described above. The starting value for the beat frequency was 50 Hz and was varied in steps of 15 Hz. The step size was reduced twice by 5 Hz after each second reversal. At the final step size of 5 Hz, six reversals were recorded and the threshold estimate was taken as the mean across these reversals. Each experimental condition was repeated four times per subject and the final threshold estimate was the mean across the four threshold estimates. In experiment 1b, the upper cutoff frequency was changed to $f_u = 1100\,\mathrm{Hz}$ in order to investigate whether the highest detectable beat frequency depends on f_u. Only condition 2 was carried out in this case. All step sizes were doubled in comparison to experiment 1a.

In experiment 2, the beat frequency was fixed to 10, 50, 75 and 100 Hz. The upper limit frequency was again set to $f_u = 550\,\mathrm{Hz}$ and only condition 2 was considered. The phase-warp stimulus $p(t)$ was mixed with binaurally uncorrelated Gaussian noise $n(t)$, similar to Siveke et al. (2007.), resulting in the mixed stimulus $s(t)$:

$$s(t) = r \cdot p(t) + (1 - r) \cdot n(t) \tag{3.6}$$

The mixing ratio r was calculated in a way that different effective, binaural modulation depths m could be generated:

$$r = \frac{1}{1+\sqrt{\frac{1}{m}-1}} \quad \text{for } 0 < m < 1; \quad r = 1 \text{ for } m = 1. \tag{3.7}$$

The adaptive parameter was the effective binaural modulation depth in dB (20 log *m*). Again, binaurally uncorrelated noise was used in the reference intervals. The step sizes in the adaptive procedure were 4, 2 and 1 dB. All other details remained unchanged in comparison to experiment 1.

3.3.2 Results

The results for the six individual thresholds and mean thresholds of experiment 1 are given in Table 3.1. The data are the mean and the standard deviation of the highest beat frequency subjects were able to distinguish from binaurally uncorrelated noise. The columns display data for the different upper cutoff frequencies f_u and for the two conditions, without and with additional high-pass noise above f_u, indicated as condition 1 and 2, respectively. For $f_u = 1100\,\text{Hz}$, only condition 2 was measured. The data show some variability across subjects with the largest deviation from the mean performance occurring for subjects MD and SE in condition 1 at $f_u = 550\,\text{Hz}$ (left-hand column). In the data for condition 2, where an additional high-pass noise was presented, the deviations between subjects are generally smaller. The mean results for the highest detectable beat frequency are 88 and 96 Hz for condition 1 and 2 at $f_u = 550\,\text{Hz}$, respectively. A one-way, repeated-measures analysis of variance (ANOVA) with the within-subjects factor condition showed no main effect of condition on the threshold data for $f_u = 550\,\text{Hz}$ [F(1,5)=1.79, p=0.24] . Since only condition 2 ensures that the subjects cannot use information from filters above the upper cutoff frequency f_u, and no significant difference between the mean results was found, all further experiments were carried out only for condition 2 which had additional high-pass noise. Comparing the results for the two different cutoff frequencies 550 and 1100 Hz (middle and right-hand column), the highest detectable beat-frequency increases roughly proportionally with the cutoff frequency. The mean thresholds for $f_u = 1100\,\text{Hz}$ is 219 Hz.

	f_u = 550 Hz (cond. 1)	f_u = 550 Hz (cond. 2)	f_u = 1100 Hz (cond. 2)
ID	66±10	85±7	173±36
MD	106±10	88±17	219±21
SE	123±16	125±9	245±28
HK	79±7	92±17	252±31
JS	72±15	91±20	228±45
FZ	84±17	95±5	196±47
mean	88±22	96±15	219±30

Table 3.1 - Individual beat frequency threshold for distinguishing phase-warp from binaurally uncorrelated noise. The values are the mean and the standard deviation from five runs. The values in the last row ("mean") are the mean and standard deviation of the individual data.

In experiment 2, six subjects participated, five of which had also participated in experiment 1. Fig. 3.2 displays the threshold modulation depth of each subject in dB for different beat frequencies of 10, 50, 75, and 100 Hz indicated on the x axis. Error bars indicate one standard deviation. The thresholds increase more or less linearly from about -7 dB (best performance) at 10 Hz to about -1 dB at 100 Hz. There was some deviation between subjects, with the most shallow decrease of performance from -5.3 dB to -1 dB from 10 to 100 Hz for subject MD and the strongest effect for subject HK (-8.2 dB at 10 Hz). On average, thresholds decreased by 1.7 dB per octave between 10 and 75 Hz. The threshold of the 100 Hz stimulus was often very close to the highest possible modulation depth (0 dB) which corresponds to the "fully modulated" phase-warp stimulus as presented in experiment 1. Four subjects (FZ, HK, BE, ID) had two or more "bad runs" in which the tracking variable reached nominally higher levels than the upper boundary of 0 dB which was allowed for the adaptive tracking variable. In this case, the run was terminated early and the respective points are shown as a triangle, indicating that the true threshold might have been higher. These data points were excluded from the mean data which is compared to the simulations in the next section. The experiment was also carried out for beat frequencies of 125 and 150 Hz where almost all threshold runs terminated early for all subjects. Thus no valid threshold values could be derived.

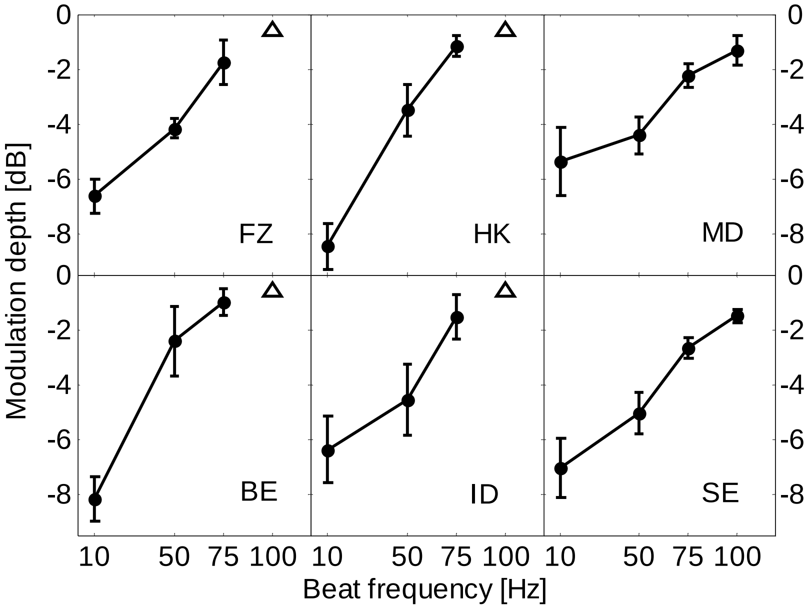

Fig. 3.2 - Modulation depth in dB at the detection threshold for different beat frequencies. For the 100 Hz beat frequency four subjects had two or more “bad runs” in which the tracking variable ran out of bounds. These points are printed with a triangle, indicating that the true threshold may be higher. Copyright Elsevier.

3.4. Simulations

This section is divided in several subsections for different types of stimuli. The emphasis is put on section 3.4.1 which contains simulations of the phase-warp stimuli that were used for the measurements. The other subsections contain simulations of standard stimuli or literature data, emphasizing the role of the modulations filters for binaural detection.

3.4.1 Phase-warp

In order to analyze the behavior of the models for the phase-warp stimuli used in the psychoacoustic experiments, only the filters with center frequencies $\leq f_u$ were considered. Both models are analyzed in the same way. As an example of the IPD model output, Fig. 3.3 shows the 450-Hz fine-structure band for a beat frequency $f_b = 4\,\mathrm{Hz}$. The cyclic variation of the rate-code based location estimate (based in turn on the IPD) is clearly visible. As described in Sec. 3.2.2 the interval [-1…1] was linearly mapped to the left-center-right location. Only small distortions by noise are observed for the low beat frequency of 4 Hz. In order to quantitatively estimate the fidelity of the model to follow the cyclic variations for different beat frequencies, the experiments from section 3 were conducted with an artificial listener.

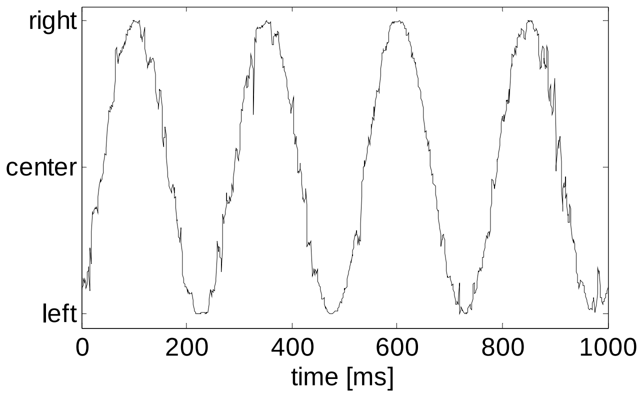

Fig. 3.3 - Simulated lateralization for a phase-warp stimulus with 4 Hz beat frequency extracted from the 450 Hz fine-structure band. Copyright Elsevier.

The identical signals and durations as in the psychophysical measurement were used except for the 20-ms ramps. The artificial listener was implemented in the same AFC framework as used for the measurements. The firing rate- and the EI_0-functions for the whole stimulus duration (1 s) of all frequency bands with $cf \leq f_u$ were Fourier analyzed for each of the presentation intervals. The artificial listener decided to take the interval with the most energy at the beat frequency component. In order to limit the model performance, an additional internal Gaussian noise, assumed at the level of the central processor, was added to the absolute value of the beat-frequency component. Without the assumption of internal noise, only the (external) noise inherent to the output of the binaural stage would limit performance. However, this inherent noise, resulting from the stochastic nature of the stimuli, would cause an excessively good model performance. The strength of the internal noise was adjusted to set the 70.7 % value of the psychometric function to the threshold beat frequency from the 550-Hz condition of experiment 1 (96 Hz). This calibration was the basis for the analysis of the 1100 Hz condition and later for experiment 2.

The 1100 Hz condition had its psychoacoustic threshold at 219±30 Hz. In contrast, the 70.7 % value of the psychometric function was found at 138 Hz for the EI model and at 159 Hz for the IPD model. In order to analyze this lack of performance, the simulations were repeated for the two models, both with and without a smoothing time constant. The results of the unsmoothed condition were 161 Hz for the EI model and 199 Hz for the IPD model.

Fig. 3.4 shows the simulations of experiment 2 along with the mean data of all subjects. Model simulations for the IPD model are given by the red dashed lines while the results for the EI model are shown in blue dotted lines. Results of simulations employing a 64-Hz filter are marked with filled circles. For both models, an additional control condition (indicated by the "x" symbols) was introduced, showing the performance without any additional smoothing ("infinity symbol"). The black circles indicate the mean experimental data for comparison. All model simulations follow the same trend as the experimental data with best performance at the lowest beat frequency (10 Hz) and increasing thresholds (worse performance) towards higher beat frequencies. While the rate of threshold increase between 10 and 50 Hz is correctly described for all models, the temporally smoothed model predictions show a degradation of threshold between 50 and 75 Hz, which is slightly too steep. Both sets of simulated thresholds are about 0.5-3 dB better than the experimental data. The unsmoothed EI model has the smallest differences with the experimental data which do not exceed 1 dB. The 3-Hz filtered EI model, represented by the "+"-symbol, only predicts a threshold value for 10 Hz with 5 dB worse performance than is observed in the experimental data.

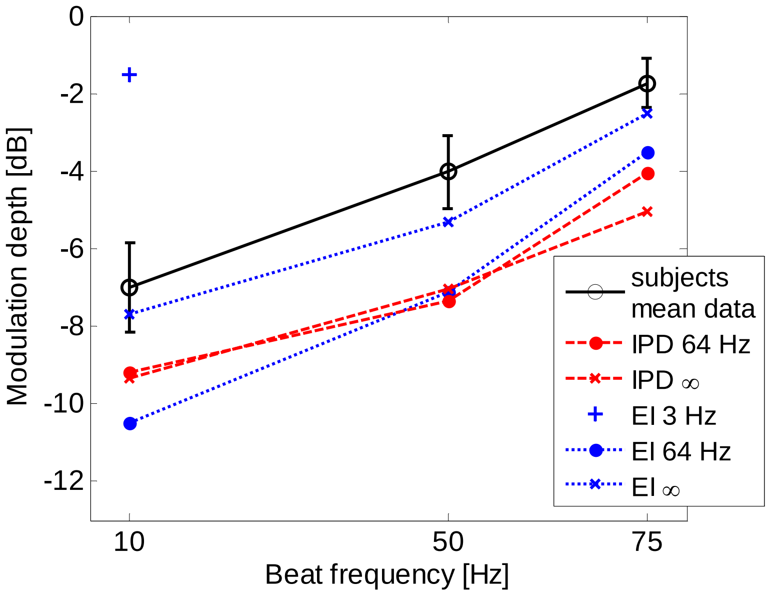

Fig. 3.4 - Mean of measured modulation depths (black circles, solid line) and simulated modulation depths in dB of the artificial listener for different beat frequencies for phase-warp stimuli with 550 Hz cutoff. IPD rate coding simulations are connected with a red dashed line, EI based simulation with a blue dotted line. The different symbols represent different low-pass characteristics of the binaural processor: "x" indicates simulations with no filtering, the filled circles represent the 64-Hz simulations, and the "+"-symbol marks the only threshold found with a 3-Hz EI simulation. Copyright Elsevier.

The frequency analysis utilized in this section implies an optimal detection of the beating, though the real spectral resolution of the binaural processor output is unknown. However, the results of the simulations do not depend critically on the spectral resolution. Comparable results were obtained with up to 19-Hz wide analysis windows which would only imply a far lower spectral resolution of the binaural detector.

3.4.2 Noise and pure tones with ITD

Figure 3.5 shows several output functions $l(t)$ of the fine-structure bands in different peripheral channels from delayed broadband noise inputs. Again, the interval [-1…1] was linearly mapped to the left-center-right location. The interaural time delay for each noise was modified in order to be exactly $\pi/2$ at the center frequency of the respective displayed peripheral channel, and the stimulus is assumed to be fully lateralized to the right-hand side. It can be seen that the lateralization is completely preserved for the peripheral channel with a center frequency of 800 Hz (upper left panel). For increasing center frequencies, the information is gradually lost until fine-structure based lateralization becomes impossible at 1.7 kHz. The output functions of delayed pure tones do not differ from the functions in Fig. 3.5 and are not plotted.

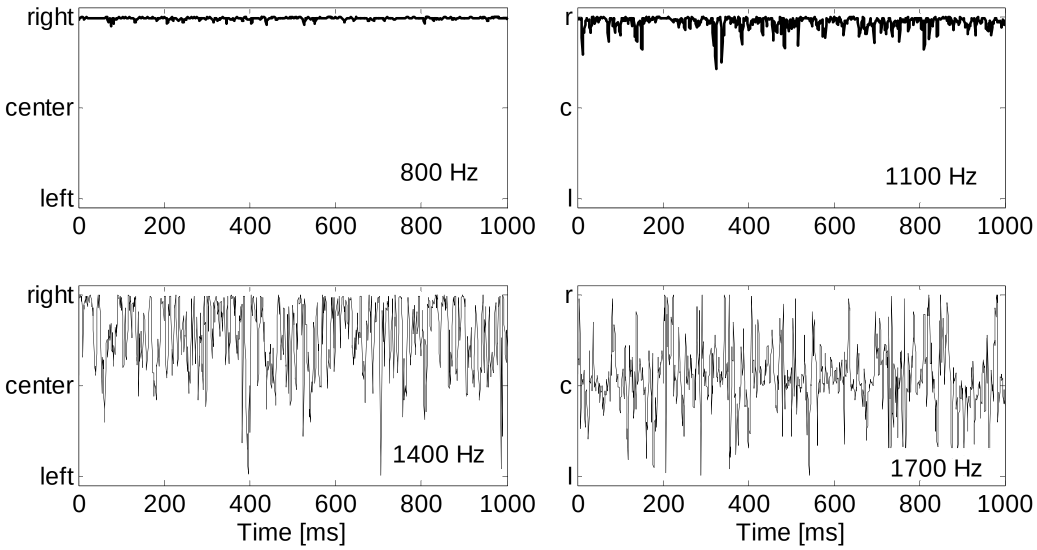

Fig. 3.5 - Model output from the fine-structure filter at different center frequencies (*cf*), indicated in the lower right corner of each panel. The signals were broadband noise stimuli with the left channel being delayed by 0.25/*cf* (IPD = π /2 at *cf*). Copyright Elsevier.

For pure tones this result is in line with psychoacoustic findings (e.g., Sayers, 1964), while sufficiently wide noise can still be lateralized even if it does not contain low-frequency components (Bernstein and Trahiotis, 1982). In this case, the information of laterality is preserved in the envelope of the noise, which is extracted by the modulation filter of the model. Fig. 3.6 displays the lateralization calculated from the 150-Hz modulation filter in the 1.7-kHz carrier band. Fluctuations in the modulation filterband are generally higher than those in the fine-structure bands with $cf < 1\,\text{kHz}$. This is related to the envelope fluctuations in the 1.7-kHz band which are generally at the lower edge of the frequency range covered by the 150-Hz modulation filter. Thus the IPD model suggests that just-noticeable differences in the envelope ITD are higher in the frequency range around 1.7 kHz than for fine-structure ITD's at low frequencies. This is in line with psychoacoustic findings from, e.g., Bernstein and Trahiotis, (1994) and Bernstein and Trahiotis (2007). Furthermore, the long cycle duration of the modulation filters can explain the laterality of very long delays (e.g., Mossop and Culling, 1998).

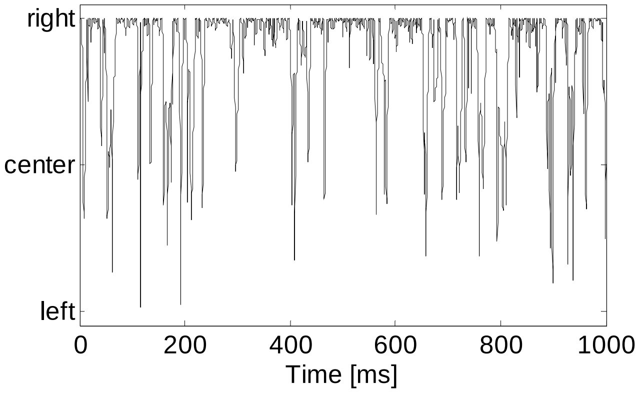

Fig. 3.6 - Model output from the 150 Hz modulation filter at 1700 Hz carrier frequency. The broadband noise input had an ITD of 0.25/150 Hz = 1.67 ms (IPD = π /2 at 150 Hz). The lateralization is mostly conserved but not as sharp as in the low-frequency fine-structure. Copyright Elsevier.

3.4.3 N_0S_π-detection

The detection of a pure tone with an IPD of π in a diotic masking noise (N_0S_π,, see Hirsh, 1948) is probably the most common test for binaural processors (e.g., Durlach, 1963, Breebaart et al., 2001b). Thresholds of this detection task can be up to 30 dB below the threshold for detecting a diotic pure tone in the same noise (N_0S_0, e.g., van de Par and Kohlrausch, 1997). The difference between the thresholds of both conditions is termed binaural masking level difference (BMLD). In the current model, the binaural processor output of diotic noise alone is always very close to zero, disturbed only by the internal noise.

The variance of the output function is close to zero. The introduction of a dichotic pure tone, however, will disturb the constant lateralization output by increasing the variance of the output function. This increase in the variance of the lateralization outputs can be interpreted as a broadening of the spatial image of the sound. Here, the assumption is that a constant lateralization output would correspond to a well focused spatial image of the sound. A fast variation of the output would correspond to a spatial broadening or smearing of the image.

In order to test whether the increased variance can be used by the IPD model as a binaural detection cue and whether it is able to predict sufficiently low pure-tone thresholds, the artificial listener was set up again. An experimental condition from van de Par and Kohlrausch (1997) was simulated. The signal tone was at 250 Hz and two bandlimited Gaussian noise maskers with a bandwidth of 10 Hz and 100 Hz were used. The 300-ms signal tone was centered in the 400-ms noise maskers. In the simulation, the same stimuli were utilized. The variance of the fine-structure lateralization output of the auditory filter tuned to the signal tone, $l^{250Hz,fine}(t)$, was determined in three consecutive 100-ms parts starting at 50 ms (at the beginning of the signal in the stimulus interval). The maximum of the three variance values was selected for each of the presentation intervals and used as decision variable in the detector. The detector selected the presentation interval with the highest maximum. In this analysis, the performance of the model was only limited by the external noise in the stimuli, an additional internal noise was not assumed. The thresholds from the literature were approximately -21 and -23 dB for the two bandwidths, respectively. These values were calculated as the difference of the N_0S_0 and BMLD values given by van de Par and Kohlrausch, 1997. The performance of the artificial listener was -26 and -29 dB, respectively.

3.4.4 Transposed stimuli

The binaural advantage simulated in Sec. 3.4.3 is gradually lost if the pure-tone frequency is increased above 1 kHz. However, van de Par and Kohlrausch (1997) created the so-called transposed stimuli, where the low-frequency signal is processed with a hair cell stage and finally multiplied on a high-frequency carrier. In this kind of stimuli, the binaural information is hidden in the envelope and can therefore not be extracted with the fine-structure filter used in 4.3. The modulation filter in the current model is directly sensitive to the envelope and can extract more or less the same information in case of a transposed stimulus as the fine-structure filter in case of the equivalent, non-transposed stimulus. When the stimuli from Sec. 3.4.3 were transposed to 4000 Hz using the technique described in van de Par and Kohlrausch (1997), the model predicted a threshold of -27 and -25 dB for the 10 and 100 Hz bandwidths using the modulation filter output variance and the same simulation procedure as in Sec. 3.4.3. In comparison, the psychoacoustic data from van de Par and Kohlrausch (1997) are about -17 and -14 dB, respectively.

3.5 Discussion

The scope of this work was to develop a model based on IPD rate coding together with realistic monaural preprocessing. The model was compared to an existing EI-type model for simulating detection thresholds of broadband binaural beat stimuli. The model was further evaluated with N_0S_π conditions and transposed stimuli and compared to literature data.

The experimental data presented in the current study provide further information on the processing of interaural temporal disparities compared to the earlier study by Siveke et al. (2007) which introduced the stimuli used here. Siveke et al. have shown that their broadband binaural-beat stimulus, referred to as phase-warp, appears to be particularly suited to investigate the basic limitations in binaural temporal resolution, since it generally leads to profoundly better detection performance for increasing beat frequencies than the common “oscillating correlation” stimulus used in earlier studies on interaural temporal resolution (Grantham, 1982; Boehnke et al., 2002; Joris et al., 2006). Thus, experimental data based on the phase-warp stimulus suggest a considerably lower estimate of a hypothetical binaural smoothing time constant (about 2.5 ms) than other studies (e.g., Culling and Summerfield, 1998; Boehnke et al., 2002). The results of the current study are in line with the findings of Siveke et al. (2007). While they used broadband stimuli covering the entire audio spectrum, the current study used bandlimited stimuli. The results of experiment 2, which used essentially the same experimental paradigm as in Siveke et al. (2007) with an upper cutoff frequency of 550 Hz, show that detection of the binaural beat is possible for rates up to about 100 Hz. Experiment 1 shows, that the highest detectable beat frequency roughly doubles to about 219 Hz when doubling the bandwidth to 1100 Hz. This result is in line with

Siveke et al. (2007) where thresholds could be derived for beat frequencies up to 256 Hz for a broadband phase-warp stimulus. The doubling of the threshold beat frequency with doubling of the stimulus bandwidth suggests that the limitations for detecting the binaural beat might not be mediated by an explicit binaural limitation of temporal acuity. Rather, it might reflect properties of the peripheral auditory system where the filter bandwidth of auditory filters limits the extraction of the binaural beats since it requires frequency components separated by the beat frequency to be coded in the converging inputs to a binaural processing stage.

The simulations of the experimental results using the model based on EI-processing and the model based on IPD rate coding show that the temporal resolution of both modeling approaches is, in principle, comparable to the resolution of the auditory system for binaural beats. In order to achieve this, the time constants for the binaural processors have to be reduced well below the 15-100 ms time constants that appear in binaural masking experiments (e.g., Kollmeier and Gilkey, 1990; Culling and Summerfield, 1998). However, higher time constants would still be necessary for modeling different masking data or other more complex tasks. In both models, this would be achieved by additional smoothing constants behind the binaural processor, which would now be assumed to be task-dependent rather than fundamental to the binaural system. The remaining 2.5 ms time constant of the binaural processor itself is comparable to the threshold period of source switching experiments (Pollack, 1978). The 2.5 ms time constant was motivated by the findings of Siveke et al. (2007). Siveke et al. fitted a smoothing time constant to their empirical data in order to compare the experimental performance in binaural modulation detection to classical monaural amplitude modulation detection. Their estimate of a 2.5 ms time constant was taken as a motivation for a smoothing filter in the IPD model. The proportional increase of the temporal resolution with f_u, suggested by the results of experiment 1, however, indicates that even a smoothing with a rather high cutoff frequency of 64-Hz (corresponding to 2.5 ms) might be overestimating limitations in the binaural system. The simulation results without the additional binaural smoothing filter in fact support the hypothesis that the intrinsic limitation caused by the filter bandwidths of the peripheral auditory system are the reason for the decrease in thresholds towards higher beat frequencies, rather than the existence of an additional explicit limitation of binaural processing. Due to the wide tails of the basilar membrane filter, the decrease in the SNR of the phase-warp beat frequency component in the model output is also rather slow, resulting in the slow degradation of model performance as a function of beat frequency, as observed in the experimental data.

Another effect of the temporal smoothing filter at 64 Hz in the models is its influence on the detection of low beat frequencies. The hypothesis is that the influence of disturbing random modulations inherent to the interaural "signal" limits the performance in the psychoacoustic measurements. In the models, investigated in the current study, there are two different filters which influence these inherent modulations. One is the temporal smoothing filter and the other is the fine-structure filter used in the IPD model. The comparison of the smoothed with the unsmoothed version showed that the unsmoothed IPD model performs a little better, while the unsmoothed EI model performs worse in experiment 2. The elevation of thresholds in the unsmoothed EI model, however, gives results that provide the best overall fit to the psychoacoustic results. Since this is a condition which includes no additional filtering (neither the smoothing filter nor the fine-structure filter) after the peripheral processing, the good agreement with the experimental data might be taken as indication that the perception is only limited by the peripheral processing and no central filtering occurs at all. However, the data of experiment 1 is best explained by the IPD model. Here the additional fine-structure filter seems necessary in order to account for the psychoacoustic data of the 1.1-kHz condition. Therefore, it seems that the psychoacoustic measures of binaural beat detection may not provide sufficient evidence to rule out one of the candidate models.

The processing based on IPD rate coding has demonstrated its ability to predict the lateralization of broadband noise and tones. It is very precise in the low-frequency peripheral channels that contain a reasonable amount of detail in the fine-structure bands. With the peripheral noise added behind the low-pass in the hair cell transduction stage, the loss of phase-locking in high-frequency channels is modeled according to physiological data of auditory nerve responses.[1]

By introducing modulation filters in parallel to the fine-structure filters after the hair cell stage, timing cues can be extracted from any carrier frequency. Furthermore, the model is also capable of qualitatively accounting for the effect of binaural masking level differences for both conventional and transposed stimuli. The model is able to use an increase in the variance of the lateralization function as a binaural detection advantage

---

[1] Models which apply the noise before the low-pass filtering still have a frequency independent SNR after the hair cell transduction stage and thus do not model the loss of phase-locking directly.

when N_0S_π is compared to N_0S_0. A threshold for the N_0S_0 condition was not explicitly simulated in the current study since the focus is the binaural processing. The N_0S_0 threshold can be well predicted by a monaural power-spectrum model of masking (Fletcher, 1940) and would be in line with experimental data for both regular and transposed conditions of van de Par and Kohlrausch (1997). The overestimation of the N_0S_π threshold or the BMLD could be reduced by the addition of internal noise or a variation of the duration of the integration time window. An extensive quantitative analysis with the necessary internal noise was not performed, since the goal was only to show a principle applicability of the model for BMLD simulations and to assess possible binaural cues for this task. A key point of the BMLD simulations is that the detector uses integration time windows of 100 ms, comparable to long integration time constants to derive the decision variable. This is the task-dependent reduction of temporal resolution which was stated earlier. The reduction is inevitable since the variance can only be determined over considerably long time intervals.

Comparing the general features of the IPD model with the EI model (as suggested by Breebaart et al. 2001a), there are two main differences: The EI model output at a specific time, and for each peripheral filterband, is a one- or two dimensional array depending on whether ILDs are considered as well. The dimensions are generally interaural delay and interaural level difference. In contrast, for each peripheral filterband, the IPD model output only consists of three scalar values: the lateralization based on the fine-structure, the envelope-based lateralization and one value for the ILD. This conceptual difference makes a direct comparison of the two models difficult. The second difference is that the IPD model operates instantaneously, using only causal filters and no internal delays. On the other hand, the processing delay of the EI model is governed by the length of the delay line, which is typically a few milliseconds.

Further investigations are required to clarify the specific benefits or drawbacks of the two modeling approaches.

3.6 Conclusions

A binaural processing model based on interaural phase differences was suggested. Comparing an established EI-type model of binaural interaction and the IPD model using the identical monaural pre-processing stages to simulate the auditory periphery, the following conclusions can be drawn:

- The IPD model can simulate experimental data on fluctuating timing disparities as a function of the fluctuation rate.
- The detection of temporally fluctuating interaural timing disparities with bandlimited stimuli in the region of 0 to 550 and 0 to 1100 Hz appears to be limited mostly by the bandwidth of the basilar-membrane filters rather than by a hypothetical sluggishness or smoothing filter in binaural processing.
- The EI-type model using established binaural integration time constant cannot simulate the results. In order to adapt the EI model to the data, the smoothing time constant of the double-exponential filter in the binaural processor output had to be reduced significantly from 30 to about 1.3 ms (according to a 3 Hz and 64 Hz cutoff frequency, respectively). With this reduced time constant, the behavior of the EI model was similar to the IPD model.
- The best match with psychoacoustic data at 550 Hz was found for the EI model predictions without any further smoothing. The data at 1100 Hz are better explained by the IPD model. Therefore, the psychoacoustic measures of binaural beat detection do not seem to provide sufficient evidence to rule out one of the candidate models.
- The second gammatone filtering used for extracting the phase in the IPD model (both for the fine-structure and the modulations) is a critical point, since this filtering is required for extracting a unique phase.
- By utilizing the fine-structure and the modulation band-pass filter, the model can account for conventional BMLD and lateralization as well as for transposed BMLD and envelope induced lateralization.

Chapter 4

Lateralization of stimuli with independent fine-structure and envelope based temporal disparities[1]

Abstract

Psychoacoustic experiments were conducted to investigate the role and interaction of fine-structure and envelope based interaural temporal disparities. A computational model for the lateralization of binaural stimuli, motivated by recent physiological findings, is suggested and evaluated against the psychoacoustic data. The model is based on the independent extraction of the interaural phase difference (IPD) from the stimulus fine-structure and envelope. Sinusoidally amplitude modulated 1-kHz tones were used in the experiments. The lateralization from either carrier (fine-structure) or modulator (envelope) IPD was matched with an interaural level difference, revealing a nearly linear dependence for both IPD types up to 135°, independent of the modulation frequency. However, if a carrier IPD was traded with an opposed modulator IPD to produce a centered sound image, a carrier IPD of 45° required the largest opposed modulator IPD. The data could be modeled assuming a population of binaural neurons with a physiological distribution of the best IPDs clustered around 45-50°. The model was also used to predict the perceived lateralization of previously published data. Subject-dependent differences in the perceptual salience of fine-structure and envelope cues, also reported previously, could be modeled by individual weighting coefficients for the two cues.

---

[1] A slightly modified version of this chapter has been published by the Journal of the Acoustical Society of America (vol. 125, p. 1622-1635, 2009) under the same title from Mathias Dietz, Stephan D. Ewert and Volker Hohmann. Copyright American Institute of Physics; with permission.

4.1 Introduction

Headphone experiments allow for an independent manipulation of interaural time differences (ITD) and interaural level differences (ILD) in a binaural stimulus. These ITD/ILD-manipulated stimuli typically result in a lateralization of the sound along an intracranial axis between both ears. The influence of an ILD is such that the intracranial image of a sound source is shifted toward the ear with the higher level (e.g., Halverson, 1922). For ILDs in excess of about 20 dB, the stimulus is perceived as shifted all the way toward one ear (e.g., Lindemann, 1986). Lateralization of stimuli with ILDs is possible over the whole audible frequency range.

Lateralization of sounds with ITDs, however, depends on the frequency composition of the stimulus. ITDs cause a lateralization of pure tones only in the frequency region below about 1.5 kHz (e.g., Kuhn, 1977), which may be related to the progressive loss of phase locking in the auditory nerve (AN) as frequency is increased (Palmer and Russell, 1986; Weiss and Rose, 1988). It is generally assumed that fine-structure ITDs are not accessible if the envelope of the signal after peripheral filtering is essentially flat and the phase of the signal, i.e. its fine-structure, is not coded on the auditory nerve (e.g., Bernstein and Trahiotis, 1996). In addition to interaural temporal disparities in the fine-structure, the auditory system is able to exploit interaural temporal disparities in the stimulus envelope (e.g., Henning, 1974; Bernstein and Trahiotis, 1985b and 1994; van de Par and Kohlrausch, 1997; Buell et al., 2008). For instance, high-frequency band-pass noise that is presented with an ITD is perceived at a lateral position even if the lowest frequency component is well above 1.5 kHz (Bernstein and Trahiotis, 1994). Sinusoidally amplitude modulated (SAM) tones have also been employed in the critical frequency region of 500 and 1000 Hz where fine-structure cues still play a role and envelope cues remain accessible (Bernstein and Trahiotis, 1985b). Either the entire waveform or only the stimulus envelope was delayed. The modulation frequency was chosen to be 50 Hz for the 500 Hz carrier and to be either 50 or 100 Hz for the 1-kHz carrier. The envelope shift alone revealed that the binaural system exploits temporal disparities in the envelope also in these low-frequency stimuli. Large differences between the results for the waveform shift and the envelope shift alone indicated a strong influence of the temporal disparities in the carrier. Furthermore, large individual differences between subjects were found.

In experiments 1 and 2 of the current paper, the individual contributions of carrier and modulator interaural phase differences[1] (IPDs) to the lateralization of SAM tones were investigated further by manipulating each of the two variables separately while holding the other fixed at 0°. Experiment 3 introduced opposing IPDs for carrier and modulator in order to determine their relative strengths, comparable to the classical time-intensity trading studies (Young and Carhart, 1974; Trahiotis and Kappauf, 1978; Ruotolo et al., 1979). Additional studies investigated the effect of modulator shape (experiment 4) and signal intensity (experiment 5).

A further goal of the current study was to develop a model of binaural lateralization that can account for the experimental data. The first conceptual model for processing interaural temporal disparities is the so-called “Jeffress model” or “delay line model” (Jeffress, 1948). The model consists of a sequence of coincidence detectors, which receive their input along two opposed chains of delay elements. In such a configuration, the ITD can be determined by the position along the delay line at which the internal delay elements compensate for the external ITD in the stimulus and the coincidence neurons at this position find the best match. Even though delay line models are able to account for a variety of interaural timing-related experimental data, the direct implementation of a delay line model (e.g., Sayers and Cherry, 1957) fails for some more complex stimuli, which carry information in both fine-structure and envelope, particularly for SAM tones (Stern and Colburn, 1978). In order to account for more complex stimuli, the delay line model has been extended in different ways (e.g., Colburn 1977; Lindemann, 1986; Zerbs, 2000; Breebaart et al., 2001a, 2001b, 2001c; Faller and Merimaa, 2004). Two influential extensions for modeling lateralization are the position-variable model (Stern and Colburn, 1978; Stern and Shear, 1996) and the weighted-image model (Stern et al., 1988). The position-variable model determines the center of gravity over the complete delay line with an optimized weighting as a function of delay. Due to the exponential decrease of the weighting for long delays, the delay line can be restricted to maximum delays of 2-3 ms. The weighted-image model determines several maxima of the cross-correlation function over a range of auditory filters and applies higher weights to those maxima which align at a constant ITD in all

---

[1] Most properties of a narrowband stimulus (e.g., tones, or wideband stimuli after peripheral filtering) can equally be described in terms of either ITD or IPD. ITD is the natural parameter to describe free field localization. However, in order to describe the model and the results of this study in the most convenient way and independent of the center frequency, the notation is mainly in terms of IPDs. Additionally the term “interaural temporal disparities” is used as a general expression that can be understood as either ITD or IPD.

auditory filters (“straightness”) and those with a generally smaller ITD (“centrality”). Finally, the weighted sum over all maxima determines the lateralization. In order to derive this weighted sum, delay lines need to be several times longer than the cycle duration of the stimulus.

A fundamental question about these models is whether or not the assumptions about the neuronal population as a function of “best delay” are in line with the populations found in physiological studies in mammals (e.g., Crow et al., 1978, Fitzpatrick et al., 1997; McAlpine et al., 2001; Brand et al., 2002; Hancock and Delgutte, 2004; Marquardt and McAlpine, 2007). The term “best delay” refers to the ITD at which a neuron has its highest response rate. Physiological findings suggest that in mammals, the best delay of a neuron is almost always within a half-cycle with respect to the frequency, where the neuron shows the highest rate response. In terms of IPD, this boundary is referred to as the π-limit (Marquardt and McAlpine, 2007). However, such a π-limit is not in line with the delay lines of 2-3 ms in the models mentioned above. By definition, a π-limited system can only detect one maximum of the cross-correlation function. If the ITD is larger than a half cycle of the respective center frequency, the detected maximum is not identical with the input ITD.

A crucial stimulus for analyzing the physiological existence of the π-limit is a noise stimulus centered at 500 Hz with a bandwidth of 400 Hz and an ITD of 1500 µs (Trahiotis and Stern, 1989). In this stimulus, the waveform is shifted by 3π/2 with respect to the center frequency which is equal to -π/2 (with the minus indicating a shift to the side of the lag). Since only -π/2 would be in the range of the π-limit, [-π; π], the highest response should switch from one hemisphere to the other when the ITD is increased from 500 to 1500 µs. This hypothesis was confirmed by Thompson et al. (2006) in an fMRI study of the human inferior colliculus (IC). They showed that the ipsilateral activity in response to the leading ear dominated in the 1500-µs condition, in contrast to a 500-µs delay where the maximum activity of most neurons is on the contralateral side. Nevertheless, the 1500-µs stimulus was still lateralized toward the “correct” leading side by all subjects (Trahiotis and Stern, 1989). Thompson et al. (2006) showed that the higher neural activity on the ipsilateral side is in line with π-limited delay lines but not with delay lines longer than 1.5 ms. They concluded that this limitation would invalidate the correlation-based models since their predictions cannot hold up with the reduced length of delay lines. The ambiguity of +3π/2 and -π/2 would force the π-limited delay line models to predict a reversed lateralization toward the lagging ear. In psychoacoustic experiments, however, this reversed lateralization is only observed for pure tones and narrowband noise with a bandwidth up to about 200 Hz (Trahiotis and Stern, 1989). Stimuli with higher bandwidths are “correctly” lateralized to the side of the lead, even though π-limited delay line models would still predict the lateralization to the side of the lag. The reason why delay line models with a π-limit cannot explain the correct lateralization is that they can no longer exploit the information in the stimulus envelope. Since envelope frequencies are much lower than the center frequency of the respective auditory filter, peaks of the envelope correlations would lie far outside of the π-limit. However, as mentioned earlier, interaural envelope cues can be assessed by the auditory system, e.g., for lateralization at high-frequencies (Bernstein and Trahiotis, 1994) or for binaural masking level differences with transposed stimuli (van de Par and Kohlrausch, 1997; Bernstein and Trahiotis, 2007).

A modeling approach that seems suitable for assessing fine-structure and envelope information, including the assumption of a π-limit, is the recent IPD model of Dietz et al. (2008). In this model, fine-structure cues and envelope cues are separated into different channels. This separation may also help in overcoming the constraints that existing hardwired models have with subject and level dependence. However, the IPD model has only been applied so far to analyze the temporal resolution of the binaural system for detecting broadband binaural beats (Siveke et al., 2007; 2008), binaural masking level differences of pure tones (Hirsh, 1948) and transposed stimuli (van de Par and Kohlrausch, 1997). For these tasks, a quantitative estimation of the perceived lateralization was unimportant and was therefore not implemented in the model.

In Sec. 4.2, an implementation of a π-limited lateralization model is developed. Thereafter psychoacoustic experiments are presented where temporal disparities of the fine-structure and the envelope cues were varied independently. In Sec. 4.4, the experimental results are compared to predictions of the lateralization model.

4.2 Model Structure

The current lateralization model is based on the IPD model as suggested in Dietz et al. (2008). In this section a more elaborate lateralization stage is introduced after a brief description of the monaural preprocessing stages:

- The middle ear transfer characteristic was approximated by a 500-Hz to 2-kHz first-order band-pass filter according to Puria et al. (1997).

- Auditory band-pass filtering on the basilar membrane was modeled with a linear, fourth-order all-pole gammatone filterbank (Patterson et al., 1987; Hohmann, 2002). Cochlea compression was modeled by an instantaneous compression with a power of 0.4 (e.g., Ruggero and Rich, 1991; Oxenham and Moore, 1994; Ewert and Dau, 2000) after band-pass filtering.

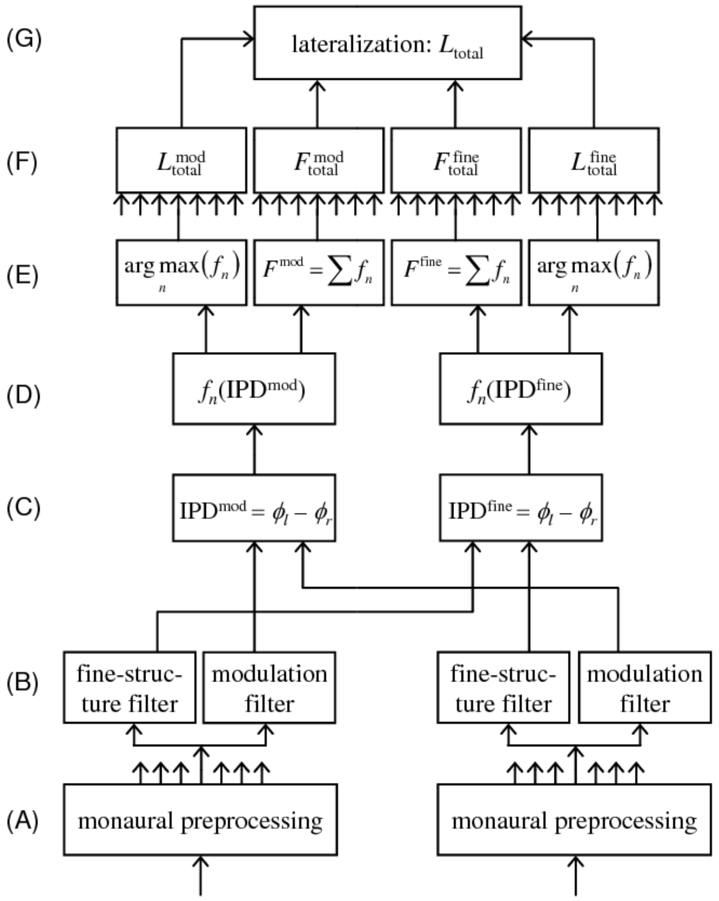
(G)
lateralization: L_{total}
(F)
$L^{\mathrm{mod}}_{\mathrm{total}}$
$F^{\mathrm{mod}}_{\mathrm{total}}$
$F^{\mathrm{fine}}_{\mathrm{total}}$
$L^{\mathrm{fine}}_{\mathrm{total}}$
(E)
$\arg\max_n (f_n)$
$F^{\mathrm{mod}} = \sum f_n$
$F^{\mathrm{fine}} = \sum f_n$
$\arg\max_n (f_n)$
(D)
$f_n(\mathrm{IPD}^{\mathrm{mod}})$
$f_n(\mathrm{IPD}^{\mathrm{fine}})$
(C)
$\mathrm{IPD}^{\mathrm{mod}} = \phi_l - \phi_r$
$\mathrm{IPD}^{\mathrm{fine}} = \phi_l - \phi_r$
(B)
fine-structure filter
modulation filter
fine-structure filter
modulation filter
(A)
monaural preprocessing
monaural preprocessing

Fig. 4.1 – Sketch of the binaural processing stages of the lateralization model. After peripheral preprocessing (output of step A), the signals are spectrally limited by two bandpass filters (step B). One filter extracts the dominant low-frequency modulation and the other extracts the fine-structure, so that the respective phases can be determined. The interaural phase differences (IPDs) are calculated (C) in order to determine the neuronal response (D). In the two outer blocks of step E, the best IPDs of the neurons with the highest response rates are determined. Furthermore, the total response is determined in the two inner blocks. Note that the two hemispheres are counteracting since the response of neurons with negative best IPDs is subtracted. Finally, the model integrates over all peripheral filter bands (F) and determines the total lateralization by a weighted spatial interpolation between the individual lateralizations $L^{\mathrm{fine}}_{total}$ and L^{mod}_{total} (G). The weighting coefficients are proportional to the two total response rates $F^{\mathrm{fine}}_{total}$ and F^{mod}_{total}.

- The mechano-electrical transduction process in the inner hair cells was accounted for by half-wave rectification with a successive 770-Hz fifth-order low-pass filter as used in Breebaart et al. (2001a).
- Uncorrelated Gaussian noise was added to all auditory bands after hair cell transformation, in order to establish a finite hearing threshold and to mimic the loss of fine-structure phase-locking for frequencies above the cutoff frequency of the filter. The noise had the same rms-value as a 0-dB-SPL, 1-kHz pure tone after half-wave rectification and prior to low-pass filtering. For the current study, however, the effect of the noise is negligible, since all stimuli were presented well above threshold.

In Fig. 4.1, the preprocessing is depicted as step (A) “monaural preprocessing”. Binaural processing starts with a separation of the monaurally preprocessed signals by a “fine-structure” and a “modulation” band-pass filter (step B). The fine-structure filter has the same center frequency as the respective peripheral auditory filter. The center frequency of the modulation (or envelope) channel was set to the frequency of the strongest envelope fluctuations. For the SAM tones used here, this is the modulation frequency of the amplitude modulation (25, 50, or 100 Hz). The use of one adjustable filter is a simplification of a modulation filterbank (Dau et al., 1997; Ewert and Dau, 2000). Both band-pass filters were realized as complex-valued all-pole gammatone filters (see Hohmann, 2002 for implementation details). The order of the filters was set to two and their equivalent rectangular bandwidth to half of the respective center frequency (i.e. $Q = 2$).

Each filter is described by the two parameters *cf* and *type*. *cf* is the center frequency of its respective peripheral auditory filter and *type* is either fine-structure or modulation. Therefore the filter outputs of step (B) can be described as functions $g_{\text{left}}^{cf,type}(t)$ and $g_{\text{right}}^{cf,type}(t)$. The IPD can now be derived from the difference of the arguments $\phi_{\text{left}}^{cf,type}(t)$ and $\phi_{\text{right}}^{cf,type}(t)$ of two corresponding filters, as shown in Fig. 4.1 (step C). A temporal smoothing of the IPD was not assumed, following the findings of Dietz et al. (2008) and Siveke et al. (2008). Furthermore, only stationary IPDs were considered in the current study and thus model predictions will not depend on the temporal resolution of the internal binaural representation of the model.

The crucial point for the lateralization is the response function of primary binaural neurons and how their response is coded in later stages of the auditory processing. For a realistic representation of lateralization, a population of binaural neurons is assumed with a characteristic tuning of their rate response to a certain best IPD (step D). The rate response function of the neurons neglects spontaneous activity and is expressed by a half-cycle of a cosine squared function:

$$f_n(\mathrm{IPD}) = \begin{cases} \cos^2\left(q\left(\mathrm{IPD} - \mathrm{BestIPD}_n\right)\right) & \text{for } \left|\mathrm{IPD} - \mathrm{BestIPD}_n\right| < \pi/(2q) \\ 0 & \text{for } \left|\mathrm{IPD} - \mathrm{BestIPD}_n\right| \geq \pi/(2q) \end{cases}, \tag{4.1}$$

with BestIPD$_n$ representing the IPD at which neuron n has the highest response rate. The parameter q determines the sharpness of the tuning around its best IPD.

The distribution of the best IPDs in terms of their frequency of occurrence in the neuron population was modeled according to data of McAlpine et al. (2001) as it is presented in Marquardt and McAlpine (2007). They recorded neurons in the inferior colliculus (IC) of guinea pigs contralateral to the leading side of the wideband noise stimulus. The best IPD distribution assumed in the present study is shown in Fig. 4.2. It was estimated from Fig. 1c in Marquardt and McAlpine (2007) by the following procedure: only BestIPD values between 0 and π (180°) were considered. The best IPDs were manually grouped in 40 bins with a width of 4.5°. 191 neurons were identified in the interval [0 180°] and used for this study. The 36 found outside this interval were discarded[1]. 7 from 234 neurons could not be identified, probably due to a strong overlap of the respective data points in the plot. The highest density was found in the two neighboring bins [45° 49.5°] and [49.5° 54°] with twelve neurons each. For IPDs between -π and 0, the distribution was mirrored and can be assigned to the IC neurons ipsilateral to the leading side of the stimulus. In terms of the model an IPD in the interval [0° 180°] means right side leading and left IC response (black bars in Fig. 4.2). An IPD in the interval [-180° 0°] means left side leading and right IC response (gray bars). It was further assumed that the IPDs in the envelope are processed in the same way as fine-structure IPDs, even though the discharge patterns of the respective neurons have some differences (e.g., Dreyer and Delgutte, 2006).

---

[1] The influence of the discarded neurons is discussed in Sec. 4.5.3

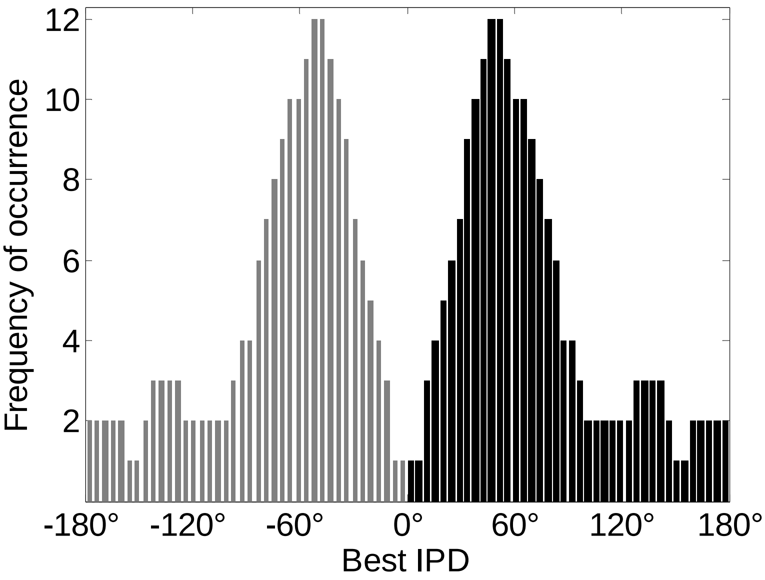

Fig. 4.2 - Distribution of the best IPDs. The data is taken from the inferior colliculus (IC) of guinea pigs (Marquardt and McAlpine, 2007). The height of each bar indicates how many neurons had their highest response (best IPD) within the respective interval. The width of each interval was set to 4.5°. An IPD in the range [0° 180°] leads to dominant response in the right IC (black bars) and an IPD [-180° 0°] to response in the left IC (gray bars).

The response functions serve for two processing steps:

First, the response functions are integrated in a way that expresses the response difference in the two hemispheres. Response from neurons of the right IC is subtracted from response of the left IC:

$$F^{cf,type} = \sum_{\text{n}|\text{BestIPD}>0} f_n^{cf,\text{type}} - \sum_{\text{n}|\text{BestIPD}<0} f_n^{cf,\text{type}} \,. \tag{4.2}$$

These quantities are displayed in Fig. 4.1 (middle blocks in step E).

Second, a place coding is employed in which the resulting lateralization estimate $L^{cf,type}$ for each specific band is determined by the best IPD of the neuron *k* with the highest response rate (outer blocks in step E):

$$k = \left(\arg\max_{n} \left(f_n^{cf,type} \right) \right) \tag{4.3}$$

$$L^{cf,type} = \text{BestIPD}_k \,. \tag{4.4}$$

Integration across different auditory bands with center frequencies *cf* is assumed as:

$$L_{total}^{type} = \frac{\sum_{cf} L^{cf,type} I^{cf,type}}{\sum_{cf} I^{cf,type}}, \tag{4.5}$$

where $I^{cf,type}$ is the intensity of the respective fine-structure or envelope channel given in dB (step F). This intensity weighted integration assumes that channels with more energy contribute greater weight to the total lateralization than channels with less energy. The general motivation of such a weighting seems intuitively plausible (further motivation is provided in the discussion). However, the specific implementation proportional to I in dB is just an assumption. Due to the middle-ear filter, the dB-SPL input is roughly transformed to dB-HL (hearing level). For high-frequency fine-structure bands, it is also damped by the hair cell low-pass filter, thus $I^{cf,type}$ represents the phase-locked intensity that is useful for binaural interaction. It was assumed that negative dB SPL values[1] of $I^{cf,type}$ are below a “binaural perception threshold” and are therefore set to zero.

The across-frequency integration reduces the estimates from each filter pair to two estimates of lateralization $L_{total}^{\mathrm{fine}}$ and L_{total}^{mod}, for the fine-structure and envelope, respectively. In the final step the overall lateralization L_{total} is determined by finding a weighted mean of $L_{total}^{\mathrm{fine}}$ and L_{total}^{mod}, based on the response $F^{cf,type}$ in each band (step F in Fig. 4.1). For such a combination no neurophysiological evidence was available. The model assumes that the response in each band is integrated over frequencies (inner blocks in step F):

$$F_{total}^{type} = p^{type} \sum_{cf} F^{cf,type} I^{cf,type}. \tag{4.6}$$

The parameter p^{type} represents the two scalar values p^{mod} and p^{fine} that can be adjusted for each subject. These parameters allow the model to account for the large individual differences found in experiments of SAM lateralization (Bernstein and Trahiotis, 1985a, 1985b) and in narrow-band noise lateralization (Trahiotis and Bernstein, 1986). The relative strength of F_{total}^{mod} and $F_{total}^{\mathrm{fine}}$ determines the weighting of $L_{total}^{\mathrm{fine}}$ and L_{total}^{mod}. These two lateralizations are the left and the right extremes of L_{total}. Now a lateralization

---

[1] Even though the energy after the monaural hair-cell processing is always above or equal to 0 dB SPL (due to the additive noise), the energy in one of the two additional filters (fine-structure or envelope) may be below 0 dB SPL. This holds, e.g., for the envelope energy in pure-tone processing or the energy in the fine-structure filter for all frequencies above about 2 kHz.

coefficient *r*, which is limited to a fixed interval, is created and the left and the right extremes are assigned to the lower and the upper limit of the interval:

$$r = \frac{F_{total}^{\mathrm{fine}} + F_{total}^{\mathrm{mod}}}{\left|F_{total}^{\mathrm{fine}}\right| + \left|F_{total}^{\mathrm{mod}}\right|}\,. \tag{4.7}$$

The denominator normalizes the difference between left and right to the interval [-1 1]. A value of $r = -1$ is assigned to the left of the two lateralizations:

$$L_{total}(r = -1) =: \min\left\{L_{total}^{\mathrm{fine}}, L_{total}^{\mathrm{mod}}\right\}. \tag{4.8}$$

Accordingly $r = +1$ is assigned to the lateralization cue which is more on the right:

$$L_{total}(r = +1) =: \max\left\{L_{total}^{\mathrm{fine}}, L_{total}^{\mathrm{mod}}\right\}. \tag{4.9}$$

Another important property of the coefficient occurs at $r = 0$: In this case, the two opposing cues are traded and the “center of gravity” of the stimulus is predicted as being perceived from the midline:

$$L_{total}(r = 0) =: 0\,. \tag{4.10}$$

Linear interpolations were assumed between these three defined values.

In the following section, psychoacoustic experiments are introduced, which serve for an evaluation of the model. In Sec. 4.4, the data of the experiments are presented and compared with predictions of the lateralization model.

4.3 Methods

4.3.1 Subjects

Four normal-hearing listeners aged between 26 and 34 years participated in the experiments. All subjects had prior experience with binaural psychoacoustic measurements. Brief training was given to the participants until they were familiar with the stimuli and the task. Subject ID received a compensation for taking part in the experiment on an hourly basis.

4.3.2 Apparatus and stimuli

The subjects were seated in a double-walled, sound-attenuating booth and listened via Sennheiser HD 580 headphones. Signal generation and presentation during the experiments were computer controlled using the AFC software package for MATLAB, developed at the University of Oldenburg. The stimuli were digitally generated at a sampling rate of 48 kHz. The transfer function of the headphones was measured in an artificial ear (B&K 4153) and digitally equalized in order to obtain a flat (± 1.5 dB) amplitude response between 0.1 and 20 kHz. The 500-ms stimuli were gated simultaneously with 20-ms raised-cosine ramps and were presented at a level of 65 dB SPL (if not otherwise stated) and with pause intervals of 300 ms.

Sinusoidally amplitude modulated (SAM) pure-tone stimuli were employed with a carrier frequency f_c = 1 kHz:

$$\begin{aligned} s_l(\mathrm{t}) &= \sin(2\pi f_c t)\cdot(1+m\cdot\sin(2\pi f_m t)) \\ s_r(\mathrm{t}) &= \sin(2\pi f_c t+\phi_c)\cdot(1+m\cdot\sin(2\pi f_m t+\phi_m)), \end{aligned} \tag{4.11}$$

where $s_l(\mathrm{t})$ and $s_r(\mathrm{t})$ represent the stimuli in the left and right ear, respectively. The phase lead of the right channel with respect to the left channel is denoted as ϕ_c for the carrier and as ϕ_m for the modulator. The modulation depth was always m = 1. The stimulus generation method allows the interaural phase of the modulator (Fig. 4.3a) and the carrier (Fig. 4.3b) to be controlled independently.

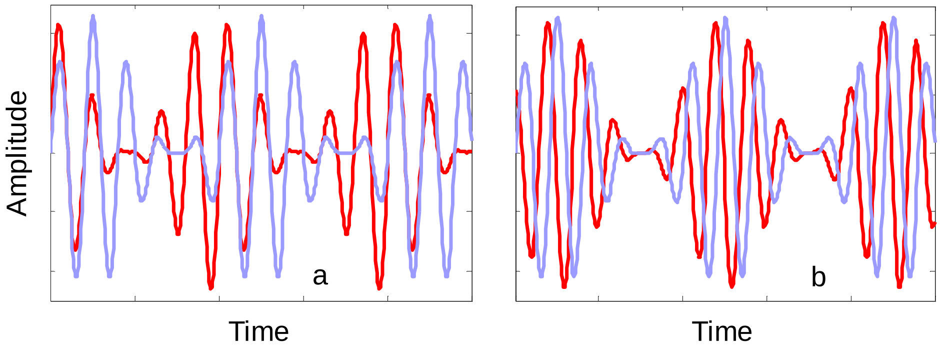

Fig. 4.3 - Two dichotic sinusoidally amplitude modulated (SAM) tones. Left channels in light blue, right channel in red. (a): Modulator shift alone. The carriers are still synchronized. (b): Carrier shift alone. The two channels have the same envelope.

In experiment 1, carrier phase shifts of 0, 45, 90, and 135° were employed. Phase shifts were restricted to 135°, since some subjects reported two intracranial positions for phase shifts $\phi_c > 135°$, one corresponding to ϕ_c and the other to $2\pi - \phi_c$ (e.g., Sayers, 1964). Modulation frequencies of 25, 50, and 100 Hz and an unmodulated condition (f_m = 0

Hz) were employed. The phase difference of the amplitude modulation was kept constant at $\phi_m = 0°$.

In experiment 2, only the modulator phase was shifted by either 0, 45, 90, or 135° and the carrier phase difference was set to zero. Modulation frequencies were again 25, 50, and 100 Hz.

In the third experiment, carrier phase differences were employed as in experiment 1 (0, 45, 90, or 135°). Additionally, a variable modulator phase difference was simultaneously applied to the stimulus. Experiment 4 was a repetition of experiment 3 with a two-tone complex (McFadden and Pasanen, 1976) instead of a SAM tone. The left channel of the stimulus was the sum of two pure tones, spectrally separated by the beat (or modulation) frequency. In the right channel, the same two tones were presented with modified phases in order to produce the desired interaural differences for the carrier and the envelope phase. The only difference from the stimulus of McFadden and Pasanen (1976) was, again, the possibility of controlling fine-structure and beat shift independently. In contrast to the “soft” sinusoidal envelope waveform of the SAM tones, the envelope waveform of the two-tone complex has steeper slopes in the envelope and shorter “modulation troughs”. In experiment 5, a simple time delay of 750 µs was applied to the entire waveform of the SAM tones (f_c = 1 kHz, f_m = 100 Hz). Therefore it was possible to investigate the relative role of fine-structure and envelope cues in a more natural matching experiment at different sound levels. These stimuli can be interpreted as the 1-kHz SAM counterparts of the 1500-µs delayed noise at 500 Hz from Trahiotis and Stern (1989), which was mentioned in the introduction. The 750-µs delay translates to a 3π/2 IPD for the carrier that can be interpreted as a π/2 IPD toward the side of the lag. The IPD of the amplitude modulation is only 27° and offers a cue toward the side of the lead. Since the sensitivity of modulation detection depends on level (e.g., Kohlrausch et al., 2000), observing lateralization as a function of level should reveal the relative importance of both cues and help in testing modeling approaches. The stimuli were therefore presented at five different sound-pressure levels: 35, 45, 55, 65, and 75 dB SPL.

3.3 Procedure

For all experiments, a two-interval, two-alternative forced-choice (2I-2AFC) paradigm was used. By pressing one of two left-right aligned buttons, the listener had to respond to the question: “Was the second stimulus left or right of the first stimulus?” In experiments 1, 2 and 5, the interaural phase shifts of either carrier (1) or modulator (2) were presented in the “target” interval according to the description in Sec. B. The perceived position was matched by an adaptively varied ILD (= $I_{\text{right}}/I_{\text{left}}$) in the reference interval. The target stimulus and the reference stimulus were randomly assigned to the two presentation intervals. When the listener’s response indicated that the position of the reference was perceived to the left of the target, the ILD was increased for the next presentation and vice versa. At the beginning of each experimental run, the ILD was randomly set to a value in the range ± 4 dB. The step size by which the ILD was initially changed was also 4 dB. After two reversals of the dependent variable, the step size was reduced to 2 dB and after two further reversals to 1 dB. The mean value of six reversals collected at the minimum step size was used as the resulting estimate. If the standard deviation of the estimate was larger than 2 dB, the run was discarded.

In experiments 3 and 4, the fixed carrier phase shift ϕ_c was combined with a modulator phase shift ϕ_m as the dependent variable. The modulator IPD was traded against ϕ_c by adjusting it to the opposing direction until the resulting stimulus was perceived from the midline. Both presentation intervals contained the same stimulus with flipped left and right channels in random order. The measurement paradigm was the same as in experiments 1 and 2, except that the step sizes for the modulator shift were $4\pi/50$, $2\pi/50$, and $\pi/50$. Trading runs with standard deviations larger than $\pi/18$ (10°) were discarded.

In all experiments the conditions were grouped in blocks. Each block had a constant modulation frequency and contained the four different IPDs in randomized order. The four blocks (experiment 1) or three blocks (experiments 2 and 3) were randomly ordered. In total this made 16 conditions for experiment 1 and twelve conditions for experiments 2 and 3. After the last block, four repetitions of all blocks were measured in the same order. Experiment 4 was performed in the same way as experiment 3 with a modulation frequency of 100 Hz only, resulting in four conditions. Experiment 5 was conducted without randomization. The five conditions resulting from the different presentation levels were measured in the order of ascending level. Five repetitions of the block were measured. For each experiment data were collected on at least two different days. At least four runs had to be valid (standard deviation smaller than 2 dB or 10°) in each condition. In the few cases in which a subject had only three valid runs an additional run was measured at the end of the experiment. The final result for each condition was the median of all valid runs.

4.4 Experimental results and model predictions

In the following, the experimental results and predictions of the model described in Sec. 4.2 are presented. For the model predictions, the individual coefficients p^{type} were chosen in units of "dB per degree". With this assumption, the model outcome is a predicted dB value according to an ILD that produces the same lateralization. The assumption implies a linear dependence between IPD and perceived lateral position in terms of ILD-equivalent as an approximation. In the matching experiments 1, 2 and 5, the target stimuli were fed into the model and the output values are directly comparable to the reference ILD in the psychoacoustic data. The output variation due to internal noise was less than 0.2 dB in most cases and was therefore ignored in the plots. The predictions for the trading experiments 3 and 4 were found by an incremental search of a sign-change in the output variable ($r \approx 0$), with increments of the modulator IPD of either 1 or 2°. For all simulations, auditory filters in the frequency range 761-1296 Hz were used. The spacing between the center frequencies of the gammatone filterbank was set to 0.2 ERB. This setting results in 10 bands below 1 kHz, one band centered at 1 kHz and 10 bands above 1 kHz. Any test with smaller spacing lead to the same results.

The parameter for the sharpness of the tuning was set to $q = 2$. This was chosen as the broadest possible tuning of the response functions that still led to meaningful predictions of the data. It is also possible to choose $q > 2$ which leads to sharper tuning of the IPD sensitive neurons.

4.4.1 Experiment 1: Matching of carrier IPD

Figure 4.4 shows psychoacoustic data (left panel) and model predictions (right panel). The perceived lateral position of the sound was matched by an ILD for the different interaural carrier phases indicated on the x-axis. The experiment was conducted without amplitude modulation (✳) and with SAM at 25, 50, and 100 Hz (indicated by □, ◇, and ○, respectively). The interaural phase of the amplitude modulation was kept constant at $\phi_m = 0$. In the model predictions (right panel), the AM had no effect, thus only a single set of data was plotted (indicated by the "+" symbols). In the left panel, each subpanel shows the median data for one subject. The error bars indicate the quartile boundaries.

An increasing carrier IPD required an increasing ILD to be matched. For all subjects except ID, there is a more or less linear relationship between IPD and matched ILD with some flattening of the curve for increasing IPDs. Subject ID required a similar ILD to match the three carrier IPDs of 45, 90 and 135°. The matched ILD range varies across subjects. The results of subject HK and subject MD differ by a factor of 2.5. The modulation frequency had no systematic effect on the results.

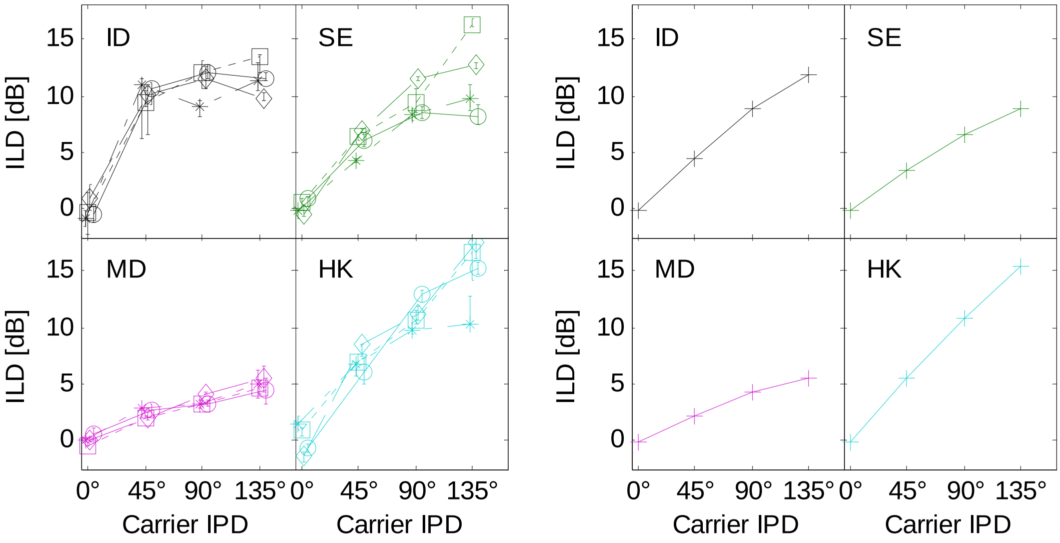

Fig. 4.4 - Left panel: Psychoacoustic results of experiment 1. An IPD in the carrier of a 1-kHz SAM tone (indicated on the x-axis) was matched with an ILD of a 1-kHz SAM tone at different modulation frequencies. The symbols indicate the modulation frequencies of ✳: 0 Hz (pure tone), ☐: 25 Hz, ◊: 50 Hz, and ○: 100 Hz. Median values are shown and the error bars indicate the quartile boundaries. Right panel: Model predictions for experiment 1. The subject dependent variable p^{fine} was fitted to the experimental data. No modulation frequency dependence was observed and only a single data set is plotted for all modulation frequency conditions, indicated by the "+" symbols.

In case of the model predictions (right panel), four individually-adjusted versions of the model are shown that account for the main features of the data. The subject-dependent constant p^{fine} was individually adjusted to give the closest fit to the data. It was set to 12.2 dB/100° for subject HK and to 10.0, 7.5, and 4.8 dB/100° for subjects ID, SE, and MD, respectively. The increase of the matched ILD is mostly linear as a result of the simplified linear transformation between angle (IPD) and ILD. A slight deviation from linearity was observed since the internal noise causes statistical fluctuations in the IPD. For IPDs of 135°, the fluctuations cause a few instantaneous IPD values >180° that are interpreted as opposing cues and reduce the mean values. Saturation effects as in subject ID for all conditions and in subject SE for the 100-Hz condition cannot be modeled with such a linear approach. The model output does not depend on the modulation frequency or on p^{mod}, since the mean value of the modulator IPD is zero in all frequency bands.

4.4.2 Experiment 2: Matching of modulator IPD

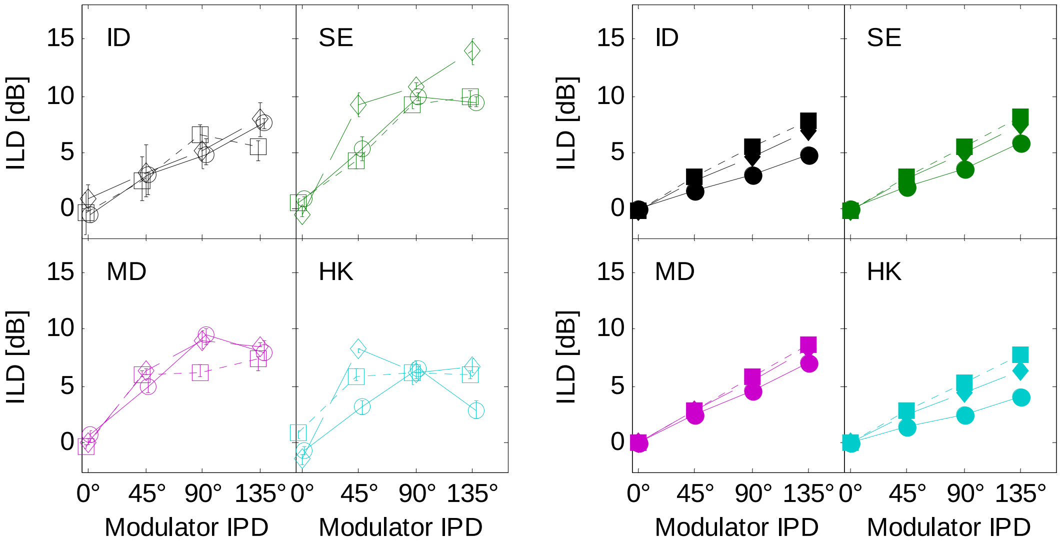

Fig. 4.5 - Left panel: Psychoacoustic results of experiment 2. Matched ILD is shown as a function of the IPD in the modulator of a 1-kHz SAM tone. Different modulation frequencies of □: 25 Hz, ◊: 50 Hz, and ○: 100 Hz were used. Median values are shown and the error bars indicate the quartile boundaries. Right panel: Model predictions. A subject independent variable p^{mod} = 7.2 dB / 100° was fitted to the experimental data and was used in all subpanels. Note that the interaural time differences introduced on the modulator are in the range of 1 to 15 ms and therefore generally much larger than interaural differences occurring in free field listening.

Figure 4.5 shows psychoacoustic data and model predictions in the same format as in Fig. 4.4. In the left panel, the matched ILDs are shown as a function of modulator IPD. Some saturation was observed for subjects MD and SE at the highest modulator IPD of 135°. Subject ID showed a reasonably linear relationship between matched ILD and IPD, while HK showed the most pronounced flattening of the matched ILD curve with increasing modulator IPD. Again, there was no systematic effect of modulation frequency on the results when the data are plotted on an IPD axis. In terms of modulator ITD, the gradient of the 100-Hz condition would be four times the gradient of the 25-Hz condition. Across-subject variability was smaller than in experiment 1. Model predictions are shown in the right panel of Fig. 4.5. The model predicts the main observations in the data. A common, subject-independent constant p^{mod} = 7.2 dB/100° was assumed. Nevertheless, the model predictions show minor dependence on subject and on the modulation frequency, most obvious in the 100-Hz condition (● symbols). The reason for this effect is a modulator induced carrier phase shift in the off-frequency bands. Further analysis is provided in the discussion. The difference between subjects is therefore caused by the different values of p^{fine} but the calibration of p^{mod} has no influence on the predictions of experiment 1.

4.4.3 Experiment 3: Trading of carrier and modulator IPD

Figure 4.6 shows the results of the trading experiment with simultaneously applied opposing interaural carrier- and modulator-phase shifts. Again, the left panel shows the experimental data while the model predictions are shown in the right panel. A peaked pattern of the trading function is observed for all subjects. A carrier IPD of 45° always required the highest modulator IPD to be traded. The resulting modulator IPDs show significant deviations across subjects. Similar to experiment 1, the largest differences are observed between subjects MD and HK who required $\phi_m = -19.3°$ and $\phi_m = -45.8°$, respectively to trade $\phi_c = 45°$ in the condition with $f_m = 100\,\mathrm{Hz}$. Unlike in experiments 1 and 2, a significant dependence of the data on the modulation frequency was observed. Larger modulation frequencies require higher opposing modulator IPDs for a midline percept. In order to define the peak position more precisely, subject MD performed an additional measurement with carrier IPDs of 30° and 60° for $f_m = 100\,\mathrm{Hz}$. The median values were $\phi_m = -14.1°$ for the trading of a 30° carrier IPD and $\phi_m = -17.2°$ for the 60° carrier compared to $\phi_m = -19.3°$ for the 45° carrier IPD in Fig. 4.6. Therefore, at least for this subject, the peak position seems to be very close to the 45° carrier IPD.

The model predictions in the right panel of Fig. 4.6 are in good agreement with the data. The peak at 45° in the simulated data is related to the high values of $F_{total}^{\mathrm{fine}}$ at this IPD which, in turn, is caused by the increased density of neurons with a best delay of 45° (Fig. 4.2). The high values of $F_{total}^{\mathrm{fine}}$ lead to a strong “reliability” of the fine-structure cue in the model. Thus the 45° carrier IPD is harder to compensate for by an opposing envelope IPD than the 90° and 135° carrier IPD, even though 90° and 135° produced a larger lateralization in experiment 1. After the p^{type} values had been adjusted to match the data in experiments 1 and 2, the model was kept unchanged for the predictions shown in Fig. 4.6. Taken together, the model accounts for the main effects in the data, i.e., the peak position, the subject dependence, and the effect of the modulation frequency.

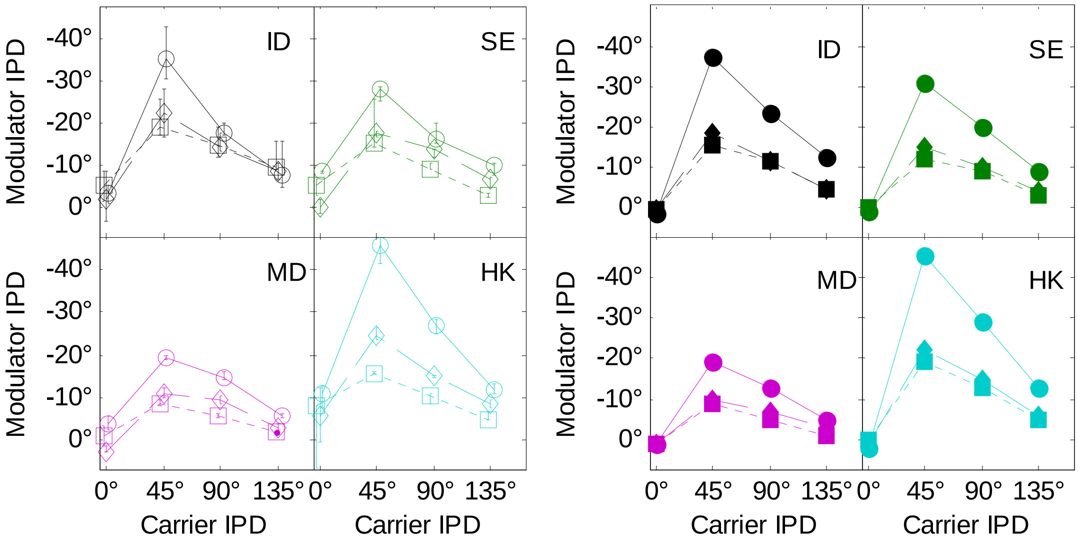

Fig. 4.6 - Left panel: Psychoacoustic results of experiment 3. The modulator IPD required for trading an opposed IPD in the fine-structure is shown as a function of the carrier IPD. The stimulus was a 1-kHz SAM tone presented at modulation frequencies of 25 Hz (☐), 50 Hz (◇), and 100 Hz (○). Median values are shown and the error bars indicate the quartile boundaries. Right panel: Trading functions predicted by the model. The parameters p^{fine} and p^{mod} were kept unchanged from experiments 1 and 2.

4.4.4 Experiment 4: Trading of carrier and modulator IPD with a two-tone complex

Figure 4.7 shows psychoacoustic data (left panel) and model predictions (right panel) for trading the envelope of a two-tone complex against a carrier IPD, comparable to experiment 3. In contrast to experiment 3 where a sinusoidal envelope waveform was used, the envelope waveform of the two-tone complex is a full-wave rectified sinusoid with the same repetition period. In psychoacoustics, the faster attack of the envelope waveform might mediate a different perceptual salience of the envelope cues (Bernstein and Trahiotis, 2007).

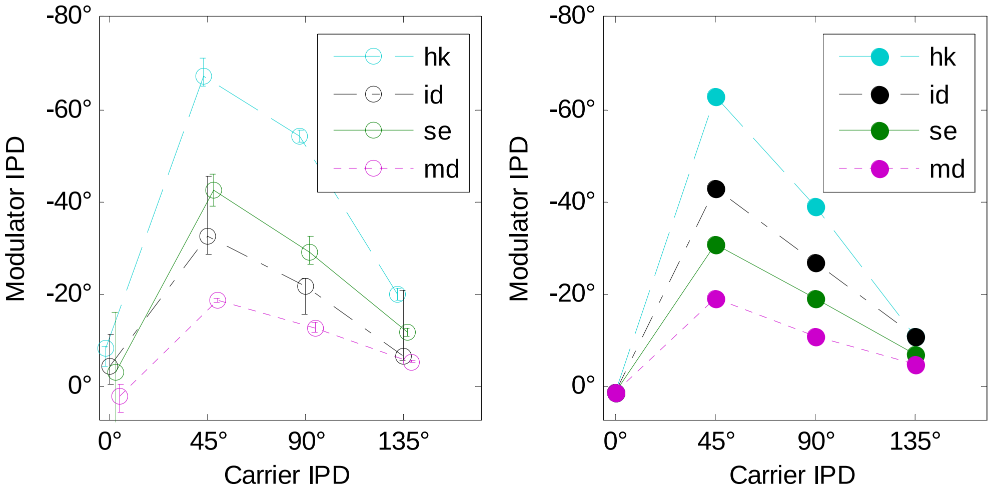

Fig. 4.7 - Trading of a two-tone complex with a beat frequency of 100 Hz using the same procedure as in experiment 3 (Fig. 6). In the left panel, the psychoacoustic trading functions of all subjects are plotted. Median values for each subject are indicated by the different symbols and the error bars indicate the quartile boundaries. The predicted trading functions are shown in the right panel. Note that in contrast to previous figures data of all subjects are plotted within one panel but with the same color coding as before.

As in Fig. 4.6, a peaked pattern is obvious in the data. In the model, the different envelope shape leads to a different ratio of fine-structure versus envelope intensity and thus to a slight increase of the necessary modulator IPD compared to the 100-Hz condition of experiment 3. A similar increase was found in the psychoacoustic data. However, in contrast to experiment 3, the values of subject SE are higher than the values of subject ID. Such an inversion cannot be modeled with fixed *p*-values for these two subjects.

4.4.5 Experiment 5: Level dependence of lateralization

The level dependence on the lateralization of SAM tones (f_m = 100 Hz, f_c = 1 kHz) was investigated, with a simple time delay of 750 µs.

Monaural modulation detection improves with stimulus level (Kohlrausch et al., 2000; Millman and Bacon, 2008). If the monaurally detected amplitude modulations form the input to a binaural processor of envelope disparities, it is likely that the envelope cues are also more salient for higher-level stimuli. In the current model this results in a level-dependent lateralization for binaural stimuli with non-zero carrier and modulator IPD. The experimental results are shown in the left panel of Fig. 4.8 for the four subjects. All subjects showed a level dependence of the matched ILD. With increasing level, the ILD required to match the fixed temporal disparities changes in the positive direction. Again, a profound across-subject variation was observed. For the lowest level, all subjects matched the perceived lateral position of the stimulus by an ILD toward the side of lag (negative ILD values). For increasing levels, the ILDs increased toward zero for ID and SE, indicating a more centered perception at the midline. For subject MD, the matched ILD switched completely from negative to positive values, indicating that the stimulus was perceived to the side of lead for high levels. For subject HK, all matched ILDs were negative, equivalent to a perception toward the side of lag. However, this subject had some differences depending on the day of the measurement indicated by the generally larger error bars.

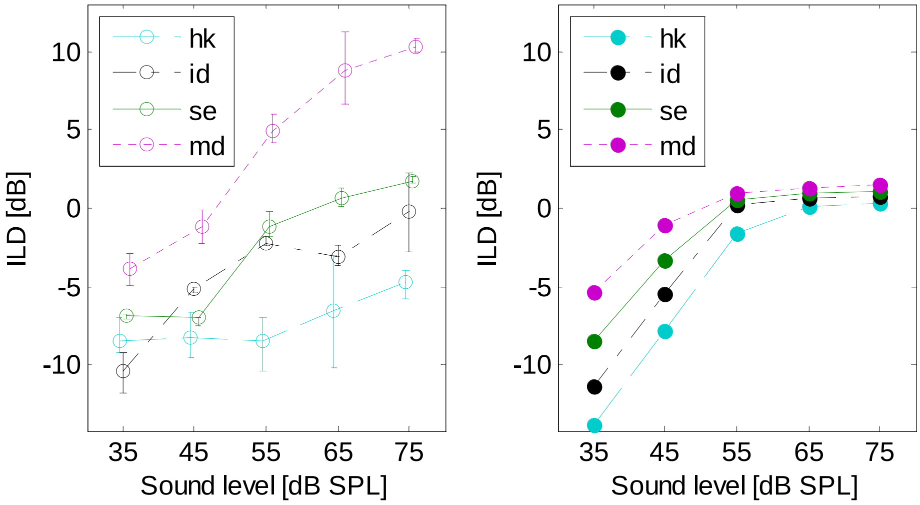

Fig. 4.8 - Left panel: Psychoacoustic results of experiment 5. Matching the lateralization of a 1-kHz SAM tone with a modulation frequency of 100 Hz, as a function of presentation level. The whole waveform was delayed by 750 µs. Median values are plotted and the error bars indicate the quartile boundaries. Right panel: Model predictions of the same experiment. Color coding is preserved from previous figures.

In the model predictions (right panel), the matched ILD changes from large negative values (equivalent to strong lateralization to the side of lag) to slightly positive ILDs (small lateralization to side of lead) as the stimulus level increases. For subject HK, the simulation predicts almost 0-dB matched ILD (centered midline perception) for the highest-level conditions. The pronounced flip to positive ILDs, as observed in the data for subject MD, cannot be explained by the model.

4.4.6 Accounting for published data

The current model was used to predict the perceived lateralization (characterized as matched ILD) of delayed noise, comparable to the stimuli described in the introduction (Trahiotis and Stern, 1989). Since both frequency region and subjects differ from the model calibration, the parameter p^{fine} was used to fit the data and p^{mod} was kept unchanged. Trahiotis and Stern (1989) reported that all of the subjects lateralized the stimuli with 50- and 100-Hz bandwidth toward the side of the lag. The 200-Hz stimulus was lateralized differently, depending on the subject, and the stimulus with a bandwidth of 400 Hz was lateralized toward the side of the lead by all subjects. These results could be modeled with $p^{\mathrm{fine}} = 1$ dB/100°: Stimuli with bandwidths below 200 Hz were predicted to be lateralized toward the side of the lag and above 200 Hz toward the side of the lead.

Another general test for the model is the lateralization of very long interaural time differences (e.g., Blodgett et al., 1956 or Mossop and Culling, 1998). For very long interaural delays, meaningful information can only be extracted from the stimulus envelope and it is therefore determined by the modulator IPD. Blodgett et al. (1956) reported that the lateralization of broadband noise can be determined correctly for delays up to 20 ms. This is very close to the longest ITDs of about 18 ms, which can be lateralized correctly by the current implementation of the model with a modulation filter at 25 Hz. Theoretically, modulation filters at lower frequencies could account for even longer delays.

4.5 Discussion

4.5.1 Psychoacoustic results

Matching the lateralization of a dichotic stimulus with an interaural level difference is a fundamental experiment in binaural psychoacoustics. Experiment 1, for instance, is comparable to the pure tone experiments of Sayers (1964). In his study, a linear dependence between ILD and IPD was found up to about 130°, where an ambiguous region began in which subjects matched the ILD either to an angle α at the right, or to $2\pi-\alpha$ at the left. Pilot experiments of this study confirmed this finding and therefore carrier IPDs were limited to a maximum of 135°, where left-right confusion occurred very infrequently. The same limit was chosen for the modulator IPDs in experiment 2. For three of the four subjects, the relation between matched ILD and carrier IPD observed in experiment 1 is linear with some flattening of the matching curve toward increasing IPDs. This is in good agreement with previously published data on pure tones (e.g., Sayers, 1964). Interestingly, the results were unaffected by a simultaneous diotic amplitude modulation (25, 50 and 100 Hz) of the 1-kHz carrier. The findings suggest that the diotic modulation offers no additional “midline cue” that might reduce the extent of lateralization of the whole stimulus.

In experiment 2, the modulator IPD was matched for stimuli with an in-phase carrier, reversing the roles of modulator and carrier when compared to experiment 1. Again, an increasing matched ILD was found for increasing IPDs with some saturation effects at 135°. The data can in principle be compared to the measurements of Bernstein and Trahiotis (1985b). They also measured isolated modulator IPDs but only up to $\phi_m = 36°$. For this modulator phase shift, the matched ILD was similar to the data of experiment 2 at 45°, except for their subject AJ, where the matching ILD was more than twice as large.

The interaction of carrier and modulator IPD was investigated in experiment 3 where both (opposed) cues were traded to obtain a central sound image. A common trading experiment for binaural cues is the time-intensity trading, where temporal (ITD/IPD) cues are traded against an opposed level (ILD) cue to perceive a midline image (Young and Carhart, 1974; Young, 1976; Trahiotis and Kappauf, 1978; Ruotolo et al., 1979). In time-intensity trading, the maximum ILD is required to trade IPDs around 90° for pure tones. If the IPD is further increased, the ILD required for trading decreases again and reaches zero for an IPD of 180°, were the IPD produces an ambiguous, semi-focused image (Young, 1976). In the current experiment 3, the pure-tone (carrier) IPD was traded by a modulator IPD instead of an ILD. A pronounced maximum in the trading function was found for a carrier IPD of 45°. This coincides with the highest neural density in the best-IPD distributions found by Marquardt and McAlpine (2007).

The purpose of experiment 4 (trading with a two-tone complex) was to investigate the influence of the envelope waveform. According to Kohlrausch et al. (2000), the modulation detection sensitivity is subject-dependent with respect to several parameters. The modulator cue in the two-tone complex might be less salient than in the SAM tone due to shorter modulation troughs in its envelope. On the other hand, the modulator cue could also be more salient because of the steeper slopes in the envelope. The data reveal that for three out of four subjects, the envelope of the two-tone complex offers indeed a less salient cue, which could be attributed to the short modulation troughs. Only subject

SE shows no decrease in the salience of the envelope cue. Transposed tones, however, offer both steep slopes and long modulation troughs and therefore cause very salient envelope cues (Bernstein and Trahiotis, 2007).

Experiment 5 revealed a level dependence of the relative importance of fine-structure and modulator IPD cues. The results show that the lateralization moves more toward the side of the lead as the sound level is increased, indicating that envelope cues become more prominent with increasing level.

4.5.2 Modeling lateralization

The suggested lateralization model is based on the physiologically motivated assumption of a π-limit for the internal, auditory representation of IPDs in different auditory bands. The assumption of the π-limit is based on the neuron distributions found by Crow et al. (1978) and Marquardt and McAlpine (2007).

Thompson et al. (2006) stated that with existing lateralization models it is not possible to explain the lateralization of 400-Hz wide 1.5-ms delayed noise centered around 500 Hz with these physiological constraints in mind. A separation of fine-structure and envelope cues for modeling lateralization with the assumption of π-limited neurons in binaural processing is required. The IPD model of Dietz et al. (2008) fulfills these requirements and was used as the basis of the current lateralization model. In order to account for the data of experiment 3, a neuron population with a physiological distribution of best IPDs had to be taken into account. An interim approach using only two pairs of IPD-sensitive neurons was not successful. In the framework of the recent discussion about multi-channel place coding versus two-channel hemisphere coding (McAlpine and Grothe, 2003; Harper and McAlpine, 2004; Phillips, 2008), the current model approach can be seen as a hybrid: The lateralization for a single channel is performed by place coding, while the combination of the channels is determined by the overall response in each hemisphere.

In comparison to other complex models of binaural processing (e.g., Breebaart et al., 2001a), the current approach is purely focused on interaural temporal disparities and cannot account for combined interaural level and temporal effects. Additionally, the process of cue combination (step E in Fig. 4.1) could be criticized as being both speculative and simplified for the purpose of trading and matching experiments. Thus, the current combination of cues, for example, does not account for an increasing lateralization in the case of congruent fine-structure and envelope IPDs. However, an increased perceived lateral position was reported with both fine-structure and envelope IPDs sharing the same direction (Bernstein and Trahiotis, 1985a). In order to account for such effects a different weighting of L_{total}^{fine} and L_{total}^{mod} would be necessary allowing for a positive summation. Independent of the weighting, a fusion to one scalar lateralization value L_{total} is always somewhat artificial as it cannot describe any spatial extent of sound images. Some listeners reported that they perceived two sound images in trading conditions and estimated a rather imaginary midpoint to perform the task. Thus, a more complete description of the spatial impression should also include a measure of image width or compactness (e.g., Hess, 2006). All necessary information to describe the spatial impression is provided by the primary binaural feature extraction of the model (e.g., Fig. 4.1 step D). It is therefore possible to model the spatial distribution by extending the “perception” stage of the model (Fig. 4.1 steps E, F, and G). In Sec. 4.4.5 the lateralization of delayed noise was simulated for different bandwidths. The model output depends on the bandwidth in the same way as the data of Trahiotis and Bernstein (1986).

However, the parameter p^{fine} had to be smaller for the delayed-noise simulations than for any of the four subjects in the SAM-tone simulations. A hypothesis for this difference is that noise shows a continuous spectrum of modulation frequencies while SAM tones have only a discrete modulation frequency. The single modulation filter did not cover all modulation frequencies of the noise; thus the influence of the modulator IPD was underestimated. In the simulation of Sec. 4.4.5 this was compensated for by an underestimation of p^{fine}. In order to exploit the whole modulation frequency range, a modulation filterbank would be necessary. This model extension is possible but not required for the SAM tones of the current study. The influence of noise bandwidth is also predicted correctly by the weighted-image model (Stern et al., 1988). However, the weighted-image model requires long delay lines of several milliseconds.

Findings of large subject variability in complex lateralization experiments (e.g., Bernstein and Trahiotis, 1985a; Trahiotis and Bernstein, 1986) can easily be modeled by the parameter p^{fine}. Conventional hardwired delay line models cannot account for these differences. In addition, the intensity weighting of the separated cues employs a level-dependence as it was also qualitatively observed in experiment 5.

4.5.3 Physiological validity of the model

A possible interpretation of the two separated fine-structure and envelope channels in the model is that in mammals the fine-structure is processed in the medial superior olive (MSO) while the modulation IPDs are processed in the lateral superior olive (LSO). There is physiological evidence for a specialization of the LSO to level and envelope (amplitude modulation) disparities (e.g., Joris and Yin, 1995 or Joris, 1996) and also several indications that the MSO is the dominant detector for fine-structure IPDs (e.g., Brand et al., 2002). In species with large heads and low-frequency hearing, like humans or dogs, the MSO is much more pronounced than in species with small heads (Grothe et al., 2001). Psychoacoustic just noticeable difference (JND) measurements of 500-Hz pure-tone and 4-kHz transposed-tone ITDs and ILDs (Furukawa, 2008) might further support the two-channel hypothesis: In transposed tones the JNDs of (envelope) ITDs and ILDs were found to add up according to a common-channel hypothesis whereas pure-tone JNDs of (fine-structure) ITD and ILD were found to add up according to partially independent channels. These two findings support the view that the envelope ITD is processed together with the ILD, while the fine-structure ITD is processed independently. On the other hand, it cannot be ruled out that cue integration works differently at low and high frequencies.

Another physiologically plausible way to explain separate access to fine-structure and envelope disparities can be obtained is the so-called DIFCOR (fine-structure) and SUMCOR (envelope) metrics (Joris, 2003). These metrics are obtained by the difference (DIFCOR) or the sum (SUMCOR) of two different cross-correlation techniques.

The distribution of best delays underlying the model was taken from Marquardt and McAlpine (2007). Due to the primary feature extraction of phase differences, the neurons that had their best delays outside the π-limit had to be discarded. Most of the discarded neurons would have been considered as “trough-type neurons” having their well-defined minima within the π-limit and two almost equally high side-peaks a half-cycle away from the minimum. This description already holds for trough-type neurons with best delays in the interval $[2\pi/3; \pi]$ (Marquardt and McAlpine, 2007). An open question is whether these neurons are really tuned to the shallow side peaks, which are separated by 2π and therefore represent the same IPD, or, more likely, to the sharp minimum which is usually close to zero IPD. The latter would support an even stricter π-limit and the respective neurons could be interpreted as excitatory-inhibitory elements (e.g., Durlach 1963; Breebaart et al., 2001a) parallel to the excitatory-excitatory majority. In the corresponding IPD region (>135°), the psychoacoustic measurements of lateralization also became instable and ambiguous and had to be excluded from the main measurements. For instance, the perception of $\mathrm{IPD}^{\mathrm{mod}} = \pi$ was described as diffuse and not lateralized (Thompson and Dau, 2008). These observations can be interpreted as caused by the side peaks of trough-type neurons and they are in line with identical activity on both ipsi- and contralateral side in the model. However, neither the model predictions nor the psychoacoustic data are distinctive enough to explain the role of peak-type vs. trough-type neurons in lateralization.

As mentioned in the introduction, Thompson et al. (2006) measured the response in the IC and not in the superior olivary complex where the MSO and the LSO are located. The 1500-µs delayed 500-Hz noise caused a higher response in the ipsilateral IC while the model output produces dominant ipsilateral fine-structure response (MSO) and contralateral modulator response (LSO). These results fit well to the MSO/LSO model interpretation since both ipsilateral MSO and contralateral LSO project excitatory to the ipsilateral IC (Loftus et al., 2004). Therefore, the model supports both conclusions of the recent study by Thompson et al. (2006): the π-limit entails a new model approach and the integrated response in contralateral vs. ipsilateral IC is not a good measure for lateralization.

A critical physiological parameter is the sharpness q of the response function. The response function was modeled as wide as possible, but with a sharpness of $q = 2$ it is still much narrower than the response functions recorded in the MSO (e.g., Brand et al., 2002). However, Fitzpatrick et al. (1997) found that subsequent neurons show a sharper tuning. Their recording from the thalamus of rabbits showed response functions that fit very well to $\cos^2\left(q\left(\mathrm{IPD} - \mathrm{BestIPD}_n\right)\right)$ with $q = 2$. Another possible fitting is a $\cos^4$ function (Harper and McAlpine, 2004) which would lead to similar results.

Overall, the most obvious and most important difference of the current model compared to other models is the separation of fine-structure and envelope cues. Psychoacoustic and physiologic findings support this approach. However, several questions remain, in particular about the role of trough-type units and the extent to which independent processing is realized in the auditory system.

4.5.4 Relation of experimental to modeled data

Reviewing the experiments, it appears surprising that the individual differences for fine-structure IPD matching are as large as a factor of 2.5. This makes it questionable to analyze mean values across subjects. An explanation for the differences in p^{fine} is that the carrier frequency of 1 kHz is already in a region where phase locking in humans decreases. Therefore the differences in the influence of the fine-structure could also be modeled by a subject-dependent variation of the phase-locking filter in the model. Currently, a cutoff frequency of 770 Hz and a 5th-order Butterworth low-pass filter is employed for all subjects. Individual differences could be modeled by varying both the cutoff frequency and the filter order or shape.

For the envelope-based lateralization, it might have been a coincidence that the variation was much less; however, most notably there was no correlation between the fine-structure factor p^{fine} and envelope factor p^{mod}, since some subjects have a larger p^{fine} and one subject has a larger p^{mod}. Therefore, it was impossible to *match* the fine-structure lateralization with the envelope-based lateralization (or vice-versa) for all subjects in the same way. On the other hand, it was possible to *trade* each fine-structure cue with an envelope cue. For instance, subject HK traded a 135° fine-structure shift with as little as 10° envelope shift to the opposite direction even though the fine-structure shift alone required an ILD of 15 dB to be matched. In the same subject, a 10° envelope IPD only required a one or two dB ILD to be matched.

While all these properties show a good accordance between model and psychoacoustic data, two points have to be discussed in some more detail.

The first point is the dependence on modulation frequency in experiment 3. For higher modulation frequencies, all subjects needed larger modulator IPDs to compensate for a given carrier IPD. One might argue that this outcome is related to the choice of an IPD-axis for the modulator and could be compensated for by converting the ordinate to an ITD-axis. However, there are several arguments against it. First, the 25-Hz trading function would be significantly higher than the two other trading functions if the modulator IPDs are converted to ITD. Since the current average factor between the 100-Hz and the 25-Hz values is 2, it will be 0.5 on an ITD-axis, which is as far away from unity than with the current IPD-axis. Second, there was no frequency dependence in experiment 2 in terms of the modulator IPD. Therefore the IPD seems to be the natural parameter for experiment 2 and it would be hard to motivate a change of the axis parameter for experiment 3. Third, the simple model hypothesis of an intensity weighting already accounts for the frequency dependence of the trading functions and can be explained in an easy way: the finite bandwidth of the auditory filters causes an attenuation of the side-bands of the SAM tones. The higher the modulation frequency is, the larger is the separation of the sidebands and the larger is the effect of attenuation, resulting in a decreased effective modulation depth. Thus, the energy in the modulation filters decreases with increasing modulation frequency and larger modulator IPDs are required to trade a given carrier IPD.

While experiments 1-3 tested different modulation frequencies, all with the sinusoidal envelope waveform and level, these features were varied in experiments 4 and 5. The reason why the model output depends on envelope waveform and level is the intensity weighting introduced to account for the modulation frequency dependence. The data from both experiments revealed the same trends but some additional variations among subjects. The variations can be explained by the nonlinear and individual increase of modulation detection sensitivity with level (Kohlrausch et al., 2000). In order to model the data more precisely, individual data on modulation detection sensitivity for each subject at each condition would be required.

The last point that needs to be discussed is the reported modulator-induced fine-structure shift in the off-frequency components. This effect, which can be seen in the simulation (Fig. 4.5, right panel), but not in the experimental data (Fig. 4.5, left panel), is caused by a modulator induced fine-structure shift in the off-frequency channels. In Fig. 4.9, it can be seen that the 90° modulator shift (f_m = 100 Hz) induces a fine-structure shift of -90° in the frequency region between 800-900 Hz. On the other hand it induces a positive shift between 1150-1250 Hz. However, the positive shift is a bit smaller and the intensity of the cue is much lower, due to the 770-Hz low-pass filter. Therefore, a net negative fine-structure cue remains, which is pulling the lateralization toward the center. The effect is stronger for higher modulation frequencies, since these stimuli result in a lower intensity at the output of the modulation filter. Furthermore, the effect in the model is stronger if a stronger fine-structure weighting is applied (e.g., to model subject HK). A possible explanation why this phenomenon is not observed in the measurements is that the auditory system does not give so much weight to off-frequency channels or that the decrease of phase-locking is not as strong as is assumed in the model. Simulations at lower carrier frequencies or with less steep low-pass filters (for example a first-order, 1-kHz low-pass as assumed in Dau et al., 1996) would also reduce the effect in the model. Another explanation would be that the auditory system has learned to compensate for this natural asymmetry.

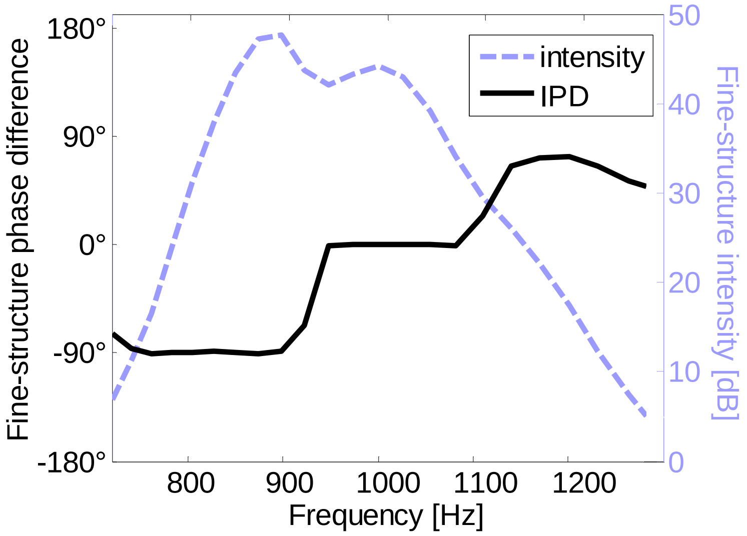

Fig. 4.9 - The two critical parameters for deriving the intensity weighted mean IPD: intensity and fine-structure IPD as a function of center frequency. The input stimulus is a 1-kHz SAM tone with a 90° shift of the 100-Hz modulator and no carrier shift. The modulator shift induces a fine-structure shift in the off-frequency channels. The lowpass filter that limits the phase-locking reduces the weighting of the high-frequency channels. Therefore the intensity weighted mean of the fine-structure IPD is negative.

4.6 Conclusions

The interaction of fine-structure and envelope interaural phase differences was investigated. Psychoacoustic data were presented and a lateralization model based on IPD extraction was suggested. In the model, the separation of fine-structure and envelope information has proven to be a successful solution under the restrictive physiological side-condition of a π-limit for IPD extraction in the different auditory filters. The following conclusions can be drawn:

1. When carrier and modulator IPDs were set in opposition, for all subjects the largest modulator IPD required to offset a carrier IPD occurred when the carrier IPD was 45°. This result could be modeled assuming a population of binaural neurons with a physiological distribution of best IPDs clustered around 45-50°.
2. Taking the neuronal population data of the distribution of best IPDs into account (π-limit), a separation of fine-structure and envelope cues appears to be required

to successfully model the lateralization of binaural stimuli independent of the psychoacoustic measurements.

3. The assumption of separate processing for fine-structure and envelope cues allows the model to mimic the strong subject- and level-dependent differences as observed in the data.
4. By assuming an effectively analog processing for fine-structure and envelope cues the influence of modulator IPDs can be modeled correctly, but only for a fixed shape of the amplitude modulation.
5. The model can account for bandwidth dependencies, since the intensity of envelope cues is influenced by the bandwidth.

Chapter 5

Model-based direction estimation of concurrent speakers from a binaural signal[1]

Abstract

Humans show a very robust ability to localize sounds in adverse conditions. Computational models of binaural sound localization and technical approaches of direction-of-arrival (DOA) estimation also show good performance, however, both their binaural feature extraction and the strategies for further analysis partly differ from what is currently known about the human auditory system. This study investigates auditory model based DOA estimation emphasizing known features and limitations of the auditory binaural processing such as (i) high temporal resolution, (ii) restricted frequency range to exploit temporal fine-structure, (iii) use of temporal envelope disparities, and (iv) a limited range to compensate for interaural delay. DOA estimation performance was investigated for up to five concurrent speakers in free field and for up to three speakers in the presence of noise or reverberation. The DOA errors in these conditions were always smaller than 5°. A condition with moving speakers was also tested and up to three moving speakers could be tracked simultaneously. Analysis of DOA performance as a function of the time resolution showed that short time constants of about 5 ms employed by the auditory model were crucial for robustness against concurrent sources.

---

[1] This chapter has been submitted for publication to Speech Communication on June 29th, 2009 in the present form by Mathias Dietz, Stephan D. Ewert and Volker Hohmann.

5.1 Introduction

Normal-hearing human listeners have a remarkable performance in estimating the direction of arrival (DOA) of specific sound sources even in the presence of a diffuse noise background, in reverberation or in the presence of other concurrent sound sources. Next to speech recognition, DOA estimation is an important part of the so-called cocktail party problem (Cherry 1953). The “problem” addresses the difficulties especially hearing-impaired listeners have in such complex acoustic environments where a specific target sound source has to be perceptually separated from the background sounds. The term “cocktail party problem” is also used in the context of computational auditory scene analysis (CASA) where performance to separate one speaker from a complex interfering background still lags behind the performance of normal-hearing humans (see e.g., Haykin and Chen 2005).

Robust machine-based DOA estimation is very important for CASA (Cooke and Brown 1993), but also for automatic clustering and segmentation of speakers in meetings (Ajmera et al., 2004) or in binaural hearing aid algorithms for self-steering beamformers (Rohdenburg et al. 2008). DOA estimation in these technical applications is usually based on broad band cross correlation in the frequency domain (Knapp and Carter 1976). More elaborate approaches of DOA estimation use frequency sub-band methods and recent studies have shown a remarkable performance of these approaches (e.g., Faller and Merimaa 2004**;** Liu et al. 2000**;** Nix and Hohmann 2006; Roman et al. 2003 and 2007; Roman and Wang 2008). These studies, as the current study, use two-channel input signals from dummy head recordings or from microphones placed close to the ears of a subject, in order to analyze basically the same information as is available to the human auditory system (except Liu et al., 2000, who did not use a dummy head between their two microphones). In the context of these studies often the term “source localization” is used instead of DOA estimation which can be somewhat misleading. For source localization, an additional estimate of the source – receiver distance would be required, which is usually not exploited in localization models. In this study the more precise term direction-of-arrival (DOA) estimation is therefore adopted.

For humans, the two most important binaural features for DOA estimation are the interaural time (or phase) difference (ITD and IPD, respectively) and the interaural level difference (ILD) (Thompson 1882). These interaural differences are suitable for DOA estimation in the frontal azimuthal plane to which most binaural models are restricted. This study is also restricted to using the IPD and ILD in frequency sub-bands.

The current study aims to adopt the strategies, processing principles and limitations of the auditory system for a human model of DOA estimation. Four specific aspects of temporal auditory processing were considered: (i) high temporal resolution, (ii) limited phase-locking range, (iii) use of temporal envelope disparities, and (iv) a limited internal ITD range. So far, these features have been considered to very different degrees in technical and auditory models. (e.g., not at all in technical approaches like Rohdenburg et al., 2008 or to some degree in auditory DOA estimation models, like limited phase-locking and a high temporal resolution by Faller and Merimaa, 2004).

An extraction of time-varying ITDs/IPD and ILD functions at a certain temporal resolution is critical for DOA estimation. Temporal integration of these functions is, however, necessary to filter out noise-induced fluctuations while a high temporal resolution is required to follow the dynamics of the interaural differences in multi-talker scenarios (e.g., Peissig and Kollmeier, 1997; Faller and Merimaa, 2004). Many psychoacoustic studies came to the conclusion that the temporal resolution of the binaural system is rather low, with time constants for temporal integration in the range of 30-200 ms (see e.g., Culling and Colburn 2000 for a review). These studies were usually focused on masking experiments and it has been questioned whether this low temporal resolution is sufficient to model the localization of concurrent speakers, where tracking of rapid changes in the interaural parameters is crucial (Peissig and Kollmeier 1997). In contrast to the masking experiments, recent studies of binaural modulation (Joris et al. 2006) or binaural beat detection (Dietz et al. 2008; Siveke et al. 2008) found a considerably higher temporal resolution of less than 5 ms. Overall, binaural temporal resolution appears to be task dependent and is probably even optimized for each task. In technical approaches of DOA estimation, temporal resolution is usually given by the block length required for processing in the frequency domain; typically 16-20 ms (e.g., Liu et al. 2000). Auditory models employ temporal integration time constants in the range of 10 ms (Faller and Merimaa, 2004) to 30 ms (Breebaart et al., 2001a). An effective method to exploit temporal integration for robust DOA estimation is a short term integration of the interaural coherence (IC). IC can be interpreted as a measure of how much of the sound arriving at both ears stems directly from a single localized sound source. As soon as a single sound source at a fixed location which would result in constant ITD/IPD and ILD functions is embedded in a reverberant environment, or is presented together with diffuse background noise or concurrent sources, the interaural coherence is decreased (e.g., Allen et al. 1977). The resulting IPD and ILD functions would be highly corrupted and a direction estimation based on these interaural parameters would not be very promising (Faller and Merimaa, 2004; Nix and Hohmann, 2006). In such conditions, the resulting listening impression is reported as “spacious” (Blauert and Lindemann 1986). Thus, the interaural coherence appears to be a good indicator for the reliability of the current interaural disparities to estimate the DOA of a directional sound source as suggested by Faller and Merimaa (2004). The human auditory system is particular sensitive to discriminate IC differences when the reference is perfectly coherent (Louage et al. 2006; Pollack and Trittipoe 1959). Goupell and Hartmann (2006) showed that the human auditory system is able to discriminate a correlation of 0.992 from a fully coherent signal, which motivates that the auditory system might be well able to employ a very sensitive coherence filter mask. The ability to detect small amounts of decorrelation makes it possible to, in principle, suppress or discard unreliable bits of the interaural functions in a time-frequency space on the basis of a reduced IC, as it was proposed also for technical applications (Allen et al. 1977; Grimm et al. 2009; Wittkop and Hohmann 2003). In this study, temporal resolution in DOA estimation tasks was assessed in combination with interaural coherence based feature selection.

The second aspect of auditory processing was the restricted frequency range in which humans can exploit temporal disparities from the stimulus fine-structure. Several psychophysical studies came to the conclusion that the ability to exploit these features starts to decrease around 1 kHz and is completely lost at about 1.5 kHz (Kuhn 1977). A physiological explanation is the reduced phase-locking of auditory nerve neurons to the stimulus fine-structure at higher frequencies (e.g., Palmer and Russell 1986). Technical DOA models do usually not consider this restriction since they have a focus on optimal performance using all information available in the signal. DOA models based on auditory preprocessing do often not consider this limitation (e.g., Supper et al. 2006; Roman and Wang, 2008), with very few exceptions like the study of Faller and Merimaa (2004). Binaural psychoacoustic models do, of course, consider this limitation in the processing stages (e.g., Breebaart et al., 2001a). However, depending on the exact implementation details, often a use of fine-structure based temporal features above 1.5 kHz was still possible. This study investigated if precise and robust DOA estimation is still possible with a strict frequency limit for the use fine-structure information as observed in human listeners.

The third aspect concerned the ability of the auditory system to exploit temporal disparities from the envelope differences between left and right ear (e.g., Bernstein and Trahiotis, 1994) in the high-frequency region, where fine-structure information cannot be exploited. Thus this third aspect was directly related to the limitation of fine-structure information discussed above. So far, temporal envelope disparities have not been directly employed in localization models or algorithms. Monaural coding of temporal envelope disparities for speech stimuli was described by Heil (2003). The question arising from this aspect was whether envelope ITDs can be extracted in an analogue way to fine-structure ITDs and how precise and robust they are for DOA estimation.

The fourth aspect took the ITD range into account that can be determined by binaural neurons. Electrophysiological data of guinea pigs (e.g., Brand et al. 2002; Marquardt and McAlpine 2007; McAlpine et al. 2001) indicated that highest response rate of binaural neurons is restricted to a maximum delay of half a cycle of the respective center frequency of each auditory frequency subband. Since the half cycle duration limit reduces the fine-structure information to the cyclic interaural phase difference (IPD), it is referred to as the π-limit (Marquardt and McAlpine 2007). Evoked potential measurements in humans (e.g., Riedel and Kollmeier 2006) similarly indicated that the neural response functions are aligned in terms of IPD rather than ITD. In contrast to these recent indications for a π-limit, localization models (e.g., Faller and Merimaa 2004, Roman and Wang 2008) and most of the psychoacoustic models (e.g., Breebaart et al. 2001a; Lindemann, 1986) assume that a mechanism is implemented in the human auditory system that can compensate for any external ITD and derive the ITD from the optimal compensation time. For humans with an ear distance of approximately 16 cm, phase leaps already occur for frequencies above 700 Hz. The influence of these phase leaps has to be investigated and strategies have to be identified that may be used to minimize the influence of this restriction on DOA estimation accuracy.

The binaural model of Dietz et al. (2008, 2009) is an auditory front-end that allows for investigating all aspects discussed above. Its standard monaural preprocessing is comparable to the preprocessing of models by Faller and Merimaa (2004) or Roman and Wang (2008). Its binaural processor extracts IPDs and is therefore π-limited. The model is further capable to extract temporal disparities from the envelope of a binaural signal. The computation of the binaural features is presented in Sec. 5.2.1 and the selection of robust estimates in Sec. 5.2.2. In Sec. 3 methods of signal generation and model training are given before the actual localization experiments are evaluated in Sec. 5.4. DOA estimation is restricted to the frontal hemisphere in order to exclude possible front-back confusions, which are not considered here.

5.2 Model structure

The DOA estimation consisted of three parts. The first part was the auditory processing for extracting the interaural parameters (Sec. 5.2.1). In the second part, the most reliable segments were extracted from the interaural functions using a coherence mask and were mapped on an azimuthal axis (Sec. 5.2.2). Part three was a task-dependent analyzer, consisting of either a temporal integrator of the short-term direction estimates from part two, or of a particle filter to track moving sources. Since the third part can be altered and was not the focus of the auditory model it is presented in the DOA evaluation section (5.3.2, 5.3.3 and 5.3.4).

5.2.1 Auditory processing

For the auditory processing the IPD model of Dietz et al. (2008) was adopted. The main processing steps are briefly summarized in the following:

- The middle ear transfer characteristic was approximated by a 500-Hz to 2-kHz first-order band-pass filter (Puria et al. 1997).
- Auditory band-pass filtering on the basilar membrane was modeled with a linear, fourth-order all-pole gammatone filterbank (Hohmann 2002; Patterson et al. 1987). The width of each filter was set to the equivalent rectangular bandwidth (ERB) of the auditory filters. 23 filter bands were implemented in the range of 200 Hz - 5.0 kHz with a spacing of 1 ERB.
- Cochlea compression was accounted for by instantaneous compression with a power of 0.4 (e.g., Ewert and Dau 2000; Ruggero and Rich 1991) after band-pass filtering.
- The mechano-electrical transduction process in the inner hair cells was accounted for by half-wave rectification with a successive 770-Hz fifth-order low-pass filter (Breebaart et al., 2001a.)

- The interaural temporal disparities were derived by band-pass filtering with a complex valued second-order gammatone filter (Dietz et al. 2008). The complex filter output

  $$g(t) = a(t) \cdot e^{i\phi(t)} \tag{5.1}$$

  contains separable information of the amplitude $a(t)$ and the signal phase $\phi(t)$. From the corresponding left-right pair of filter outputs, g_l and g_r, the interaural transfer function (ITF) was calculated:

  $$\mathrm{ITF}(t) = g_l(t) \cdot \overline{g_r}(t) = a_l(t) \cdot a_r(t) \cdot e^{i(\phi_l(t) - \phi_r(t))} \ . \tag{5.2}$$

  The ITF is still complex valued and contains both phase and amplitude information. It is therefore ideal for temporal smoothing of the interaural functions. A temporally smoothed IPD was then extracted from the low-pass filtered ITF:

  $$\mathrm{IPD}(t) = \arg\left(\left[\mathrm{ITF}(t)\right]_{lp}\right), \tag{5.3}$$

  with “*lp*” denoting a low-pass filter which will be investigated in Sec 4.1. The IPD can optionally be translated to an ITD through division by the instantaneous frequency

  $$f_{inst}(t) = \frac{1}{2 \cdot 2\pi}\left(\frac{\mathrm{d}\phi_l(t)}{\mathrm{d}t} + \frac{\mathrm{d}\phi_r(t)}{\mathrm{d}t}\right). \tag{5.4}$$

  The complex valued filter allows for deriving the ITF and thus the IPD of the temporal fine-structure or the envelope of the signal. The first was achieved by centering the filter at the same frequency as the preceding auditory filter. For the envelope IPD, the filter was centered at a modulation frequency of interest, preferably at the output of a high-frequency auditory filter.

- In addition to the model published in Dietz et al. (2008), a second-order low-pass modulation filter with a 30-Hz cut-off frequency (operating in parallel to the fine-structure and modulation band-pass filters) was introduced. The modulation low-pass filter was employed to derive the ILD from the energy ratio of the two low-pass filter outputs h_r and h_l:

  $$\mathrm{ILD}(t) = \frac{20}{c} \cdot \log_{10}\left(\frac{|h_r(t)|}{|h_l(t)|}\right). \tag{5.5}$$

The ILD was expressed in dB and was divided by the compression exponent c in order to scale the internal representation to the original ILD occurring at the ears prior to basilar membrane compression.

A sketch of the current model processing is shown in Fig. 5.1. This model is very restrictive in terms of mechano-electrical transduction and in terms of deriving the interaural temporal disparities. The primary IPD extraction directly includes the π-limit.

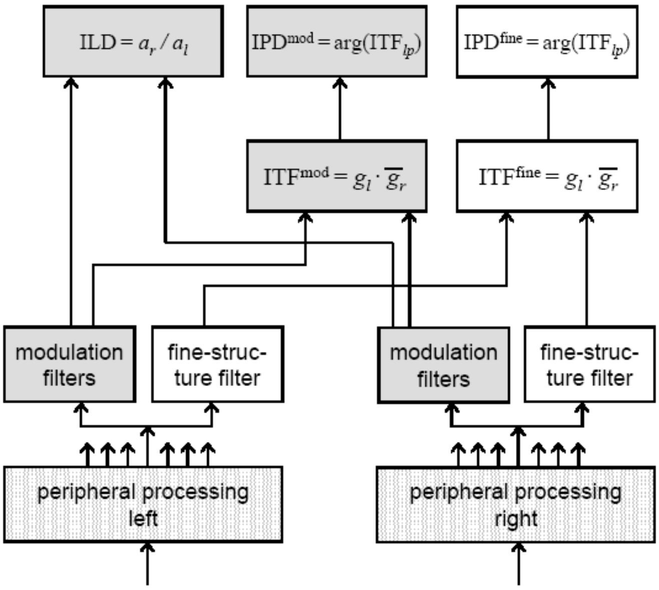
$\mathrm{ILD} = a_r / a_l$
$\mathrm{IPD}^{\mathrm{mod}} = \arg(\mathrm{ITF}_{lp})$
$\mathrm{IPD}^{\mathrm{fine}} = \arg(\mathrm{ITF}_{lp})$
$\mathrm{ITF}^{\mathrm{mod}} = g_l \cdot \overline{g_r}$
$\mathrm{ITF}^{\mathrm{fine}} = g_l \cdot \overline{g_r}$
modulation filters
fine-struc-ture filter
modulation filters
fine-struc-ture filter
peripheral processing left
peripheral processing right

Fig. 5.1 - Processing stages of the auditory model. Peripheral processing splits the input signal in 23 auditory filters per ear, followed by half-wave rectification, low-pass filtering and compression. Only one of these channels is drawn for the further processing blocks. The IPD is derived for the fine-structure and envelope. Additionally the ILD is derived at the output of an envelope low-pass filter. The model architecture was taken from Dietz et al. (2008) and was extended by the ILD processing.

An advantage of the current model is the high temporal resolution of the interaural functions which can be derived on a sample by sample basis. Another convenience is the additional gammatone filtering following the mechano-electrical transduction model. At low frequency channels up to about 1.4 kHz, the main reason of these filters is to separate DC-components from the temporal fine-structure as it is required for the phase calculation. At higher frequency channels, where no temporal fine-structure information is left, parallel filters can be applied in the form of a modulation filterbank, or filters can be set or adapted to, e.g., the fundamental frequency or pitch region of a target speaker. In order to set the filters to the correct pitch region, additional pitch estimation was necessary which was performed with the YIN algorithm (de Cheveigné and Kawahara 1999; de Cheveigné and Kawahara 2002).

5.2.2 Feature selection

As mentioned in the introduction, a coherence filter makes is possible to discard “corrupted” segments from the interaural functions which are not likely to originate from a point source (Allen et al., 1977; Faller and Merimaa, 2004). Recently also ratio and binary time-frequency masks which are closely related to optimal Wiener filters have been developed for binaural speech recognition and separation (Roman et al., 2003; Srinivasan et al., 2006; Li and Wang, 2009). So far the psychoacoustic justification for these masks has not been evaluated. These masks were therefore not considered for the realistic auditory feature extraction in this study.

Since the binaural processor used in this study does not rely on cross-correlation, the interaural coherence (IC) was not directly assessable. However, Goupell and Hartmann (2006) have shown that the temporal fluctuations of the interaural functions are possibly an even better measure for psychoacoustic decorrelation sensitivity. In the current model, the IPD fluctuation is directly accessible and was specified in the form of the interaural vector strength, $\mathrm{IVS_G}$:

$$\mathrm{IVS_G}(t) = \frac{1}{\tau_s} \cdot \left| \int_0^\infty \mathrm{d}\tau e^{i \cdot \mathrm{IPD}(t-\tau)} e^{-\tau/\tau_s} \right|, \tag{5.6}$$

where τ_s is a time constant for the temporal integration. This approach has been used as a “coherence filter” in Grimm et al. (2009) and originates from a more general expression by Wittkop and Hohmann (2003). The IC and $\mathrm{IVS_G}$ were compared for a concurrent speaker situation. Both measures were found to be are almost identical and the choice of the filter mask type had no influence on the DOA estimation performance of the current model. Another measure similar to the vector-strength is

$$\mathrm{IVS}(t) = \frac{\left| \int_0^\infty \mathrm{d}\tau \mathrm{ITF}(t-\tau) e^{-\tau/\tau_s} \right|}{\int_0^\infty \mathrm{d}\tau \left| \mathrm{ITF}(t-\tau) \right| e^{-\tau/\tau_s}}, \tag{5.7}$$

where ITF denotes the interaural transfer function introduced in Sec. 5.2.1. The difference between IVS and the definition $\mathrm{IVS_G}$ is the intensity weighting which was discarded in $\mathrm{IVS_G}$ by using the IPD instead of the ITF. The intensity weighting allows the IVS to reach high values after a source-onset faster than the $\mathrm{IVS_G}$. Therefore the

IVS was used to derive a filter mask which consists of a binary weighting w_1 of the interaural parameters based on a threshold value IVS_0 for the interaural vector strength:

$$w_1 = \begin{cases} 1 & \text{if } \mathrm{IVS} \geq \mathrm{IVS}_0 \\ 0 & \text{if } \mathrm{IVS} < \mathrm{IVS}_0 \end{cases}. \tag{5.8}$$

The temporal integration time constant τ_s was set to a multiple of the cycle duration T_c corresponding to the center frequency f_c of the respective auditory or modulation filter. This variable smoothing time constant resulted in frequency independent convergence of the IVS functions, so that it was possible to apply a frequency independent IVS threshold. This is in contrast to Faller and Merimaa (2004) who used fixed time constants. A third novelty of the mask used in this study was the additional binary weighting w_2:

$$w_2 = \begin{cases} 1 & \text{if } \dfrac{\mathrm{d}\,\mathrm{IVS}(t)}{\mathrm{d}t} \geq 0 \\ 0 & \text{else} \end{cases}. \tag{5.9}$$

The reason for this additional “rising slope condition” was that the onset of an interferer led to immediately corrupted binaural parameters, while both IC and IVS needed a finite time to drop below threshold. These short but misleading time segments could easily be filtered out by applying the above slope condition.

A general problem of the interaural parameters IPD and ILD is their ambiguity with respect to the azimuth angle α. The ILD shows a maximum at about 60°. Therefore an ILD of 90° is identical to a certain ILD < 60°. A further disadvantage of the ILD is that the ILD for a certain target source can be “dragged” towards the ILD of an interferer source (see e.g., Nix and Hohmann, 2006). The suitability of the ILD alone is therefore very limited for quantitative direction estimation. For the IPD, on the other hand, the physiologic limitation of the neural tuning leads to mapping ambiguities if the cycle duration of the auditory filter output is shorter than the ITD range that can occur for a human model subject (physiologic range: about -700 to +700 µs). Therefore the IPD becomes ambiguous above about 700 Hz even though psychophysics indicates that it can be exploited up to about 1.4 or 1.5 kHz. In headphone experiments IPD ambiguities have been studied (e.g., Sayers 1964). For instance pure tones with a large IPD (> 135°) led to a percept of hearing the tone sometimes from the left, sometimes from the right or from both positions simultaneously (Sayers 1964). However, with an additional small ILD the tone on the slightly louder side quickly dominated the perception, even if the

IPD alone favored the other direction (Dietz et al. 2009). It is therefore reasonable to assume that the IPD allows for two possible azimuth direction percepts: the first at $\alpha_1 = p_f\left(|\mathrm{IPD}|\right)$ and a second at $\alpha_2 = p_f\left(|2\pi - \mathrm{IPD}|\right)$, with p_f referring to an arbitrary IPD-to-azimuth-direction mapping function. In conjunction with these two directions, the sign of the ILD which has no ambiguity for azimuth angles can decide for the "correct" percept. This psychoacoustic finding was implemented by setting the ambiguous sign of the IPD to the more reliable sign of the ILD. In this way, the qualitative direction estimation was performed on the basis of the IPD with ambiguities resolved by the ILD sign (and discarding the unreliable size of the ILD).

Figure 5.2 shows how this ILD-controlled IPD unwrapping extends the unambiguous IPD region from [-π; π] to [-2π; 2π]. Correspondingly, the frequency range for IPD-to-azimuth mapping is extended from about 700 Hz to 1400-Hz. In order to avoid erroneous unwrapping for small α, where IPD and ILD are both close to zero but differ in sign, unwrapping was only performed for ILD values equal or above 2.5 dB. The IPDs extracted from the modulation filters in high-frequency channels showed no ambiguities, since typical modulation or pitch frequencies were well below 700 Hz.

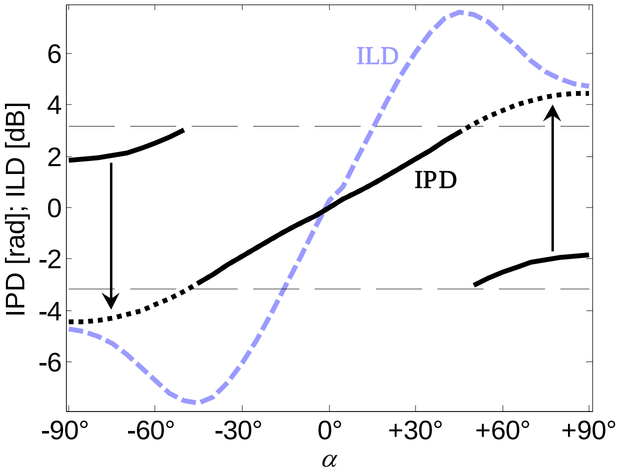

Fig. 5.2 - Median fine-structure IPD and ILD in the 1-kHz band as a function of azimuth for a single undisturbed speaker in the free-field condition. Phase leaps of the IPD and non-monotonic behavior of the ILD do not allow for direct azimuth mapping. The phase leaps can be avoided by setting the IPD sign to the unambiguous sign of the ILD. The resulting unwrapped IPD is injective up to 1.4 kHz and can be employed as input to an IPD-to-azimuth mapping function for all fine-structure information.

In order to perform the DOA estimation, the model had to "learn" which angle corresponded to which unwrapped IPD. To establish the IPD-to-azimuth mapping function, IPDs were extracted for a range of known azimuth angles resulting in a discrete function IPD(α,f), where α denotes the azimuth ($\Delta\alpha$ = 5°) and f the center frequency of the respective auditory filter. The derivation of the IPD-to-azimuth mapping function is described in Sec. 5.3.2.

5.3 Method

5.3.1 Signal generation

The spatial binaural audio inputs were generated from "dry" and "clean" sound source signals by means of virtual acoustics. The signals at the both ears were gained by convolution of a single-channel source signal with head-related impulse responses (HRIR) of the left and the right ear for a specific spatial position of the source. In the case of several superimposed sources, the ear signals for each spatial source were added up.

The HRIRs were taken from a database recorded in the Oldenburg lab with a Brüel & Kjær Type 4128C head and torso simulator by Kayser et al. (2009). If not otherwise stated, free-field HRIRs were employed with a source in 3 m distance and with an elevation angle of 0°. The microphones were placed in the ear canal. With a spacing of 5° in the azimuthal direction, the subset of HRIRs used in this study contains 72 HRIRs. When moving speakers were simulated, neighboring HRIRs were interpolated. The interpolation algorithm separates the complex spectra in a real and an imaginary part and interpolates both parts linearly.

In order to investigate the influence of reverberation, a binaural room impulse response (BRIR) of an office room was taken from the database. The speaker was positioned at 45° to the dummy head. For these measurements only impulse responses recorded with hearing-aid microphones behind the ear were available.

The source signals that were simulated to be emitted by the spatial sources in the virtual acoustics setup were speech signals from the TIMIT database (Garofolo et al. 1990) containing single sentences in English. For a superposition of concurrent speakers, all sentences were uttered by different speakers, started simultaneously and terminated when the first speaker finished the sentence. All speakers had the same presentation level.

If noise was added to a virtual acoustic scene, omnidirectional speech-shaped noise was used. This noise was generated by superimposing different speech-shaped noises from each direction in the (full) horizontal plane. Each speech-shaped noise was generated by adding 100 TIMIT sentences with randomized starting positions.

5.3.2 IPD-to-azimuth-direction mapping function

In order to estimate a sound-source direction from the IPD, a mapping function from the unwrapped IPD (IPD_u) to an azimuth direction had to be established. For this a short set of 10 speech segments of 2.7 s each was convolved with each of the 37 labeled HRIR from the anechoic chamber (frontal hemisphere). The resulting ear signals (without reverberation, other disturbing sources or noise) led to very precise and robust interaural parameters. The median of IPD_u for each of the 37 positions was derived for each auditory filter and stored in a matrix. Since a continuous mapping for any input IPD_u was necessary, a 9th-order polynomial p_f was fitted to the 37 values for each fine-structure and modulation filter with center frequency f. For the DOA estimation, the azimuth direction estimate α was then derived by the resulting IPD-to-azimuth mapping function:

$$\alpha = p_f\left(\mathrm{IPD}_\mathrm{u}\right). \tag{5.10}$$

5.3.3 Across filter integration

The optimal strategy of deriving the source direction from the azimuth estimates in the different fine-structure and modulation filters of the model might be task dependent. A human listener might focus on a specific filter or frequency region with a good signal-to-interferer ratio. In other cases, the listener might simply integrate the information across all filters. The following simplified procedure was employed here:

Twelve fine-structure bands were used in the frequency range from 200 – 1400 Hz with a spacing of the equivalent rectangular bandwidth (ERB). The direction estimates from all fine-structure filter bands were pooled to derive a single fine-structure-based estimate. The estimates from the modulation filters were treated separately and were divided into two sub groups. The first group was based on modulation filters which extract information from auditory filters below the 1.4 kHz fine-structure limit, working in parallel to the fine-structure filters. The second group was based on modulation filters which extract information from auditory filters above the 1.4 kHz fine-structure limit. Taking into account direction estimates from the first modulation-filter group did never increase the overall performance when combined with the respective fine-structure-based estimates. These modulation-filter-based estimates were therefore discarded in this study. The results of the 11 modulation filters in the high-frequency bands (1.4 - 5.0 kHz) were pooled to obtain a second, envelope-based direction estimate in addition to the fine-structure-based estimate. The modulation center frequency was kept constant for all 11 filters. Since the pitch of the speakers was derived before, a group of modulation filters was generated for each speaker with the modulation frequency equal to its median pitch. Depending on the evaluation task (as described below), the envelope-based direction estimate was taken into account.

5.3.4 Task-dependent direction estimation

The sequence of azimuth estimates as a function of time which was available after applying the IVS mask served as the input to the direction estimator. However, the optimal strategy of deriving possible source directions from the data in arbitrary auditory scenes depended on the scene itself. The focus can be either on the direction of static sound sources which are either speaking simultaneously or sequentially or the focus can be on the tracking of moving sound sources. A-priori knowledge may or may not be available to the analyzer such as e.g. the number of sources or the spectral and temporal dynamics of a source. The direction estimation in this study only used the median fundamental frequency of each speaker to set the modulation filters and used some a-priori knowledge about the motion dynamics, e.g. stationarity. Two standard problems were investigated exemplarily: Direction estimation of concurrent non-moving speakers and tracking of concurrent moving speakers.

For non-moving speakers, the azimuth estimates were collected for the whole observation time of the scene in an azimuth histogram. A Gaussian mixture model (GMM) was applied to the azimuth histogram to extract direction estimates as well as the number of active speakers. Some a-priori knowledge was necessary for the Gaussian mixture model, since the number of Gaussians had to be larger or equal to the number of sources. Since the number of sources never exceeded five concurrent speakers in this study, the number of Gaussians was set to seven somewhat arbitrarily.

For tracking the direction of moving speakers, different methods such as particle filters (Nix and Hohmann 2007) or Hidden Markov Models (Roman and Wang 2008) have been applied successfully. In order to apply any of these methods, reliable direction estimates from observation intervals less or equal to the motion time scale of the sources are required. The focus is therefore put on the required observation time for reliable DOA estimation of simultaneous speech sources rather than on the details of the statistical tracking method. An example is given for the tracking of concurrent moving speakers with a previously published model for multiple source tracking based on Rao-Blackwellized particle filters (Särkkä et al. 2007).

5.4 Model evaluations

5.4.1 Temporal resolution

In the first evaluation condition, two to five stationary speakers were localized simultaneously. The model performance was evaluated for several integration time constants of the IVS filter mask. The integration time constant that led to the best results in this condition was then used for the other evaluation conditions.

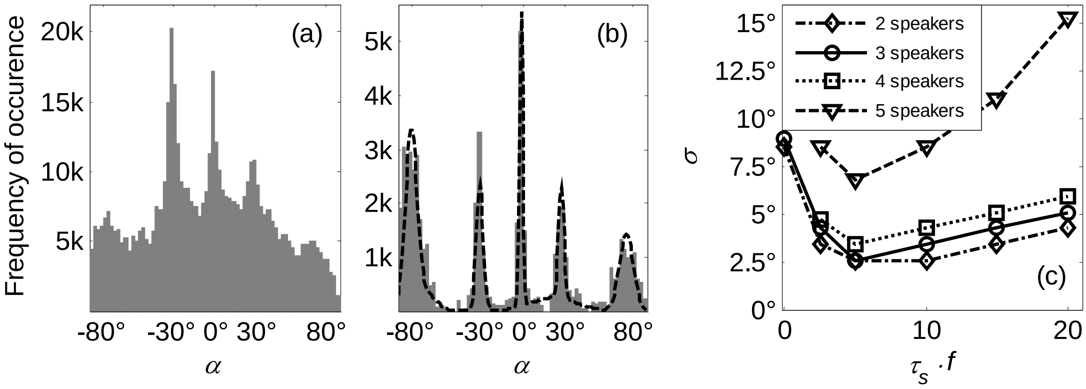

Fig. 5.3 - Panel (a): Pooled histogram of the DOA estimates from all 12 fine-structure filters. Five stationary speakers at α = -80°, -30°, 0°, +30° and +80° were observed for 3.3 s. Panel (b): same condition as (a) but only for DOA estimates for instants in time that met the IVS > 0.98 criterion. The IVS integration-time constant was $\tau_s = 5 \cdot T_c$. The dashed line in panel (b) indicates the result of the GMM. Note the different y-axes scaling for (a) and (b). Panel (c): Standard deviation of the peak at -80° of the GMM output for different integration times and different numbers of speakers. Speaker directions as in (a) were added subsequently for azimuth angles from left to right.

A target speaker from -80° was superposed by an increasing number of 1 to 4 concurrent speakers. The concurrent speakers from azimuth directions of -30°, 0°, +30° and +80° were subsequently added with each having the same sound-pressure level as the target. For each number of sources, the azimuth estimates derived for a 3.3 s observation interval (duration of the sentences) were collected in a histogram with 2.5° bin width. A Gaussian mixture model with seven Gaussians was used in order to fit the distribution. Exemplary histograms are shown in Fig. 5.3a without and in Fig. 5.3b with additional IVS filtering for the condition with 5 superposed speech signals. The variance of the Gaussian fit to the histogram resulting from the -80° speaker was evaluated as a measure of the precision of the direction estimate. For each of the four conditions with 1-4 concurrent speakers (2 to 5 speakers overall), five different integration time constants τ_s (2.5, 5, 10, 15 and 20 cycles) for the IVS mask and no temporal integration (0 cycles) were tested, resulting in 24 conditions. The resulting variances in the case of a successful direction estimate are shown in Fig. 5.3c. A condition was defined as "successful" if the number of dominant peaks is equal to the number of input speakers and all estimated directions differ by less than 10° from the input direction. It is observed that the performance decreases with increasing number of concurrent speakers. For four and five speakers it was not possible to successfully derive a direction estimate without an IVS mask ($\tau_s \cdot f = 0$ in Fig. 5.3c).

The 20 conditions with temporal integration and IVS masking ($\tau_s \cdot f > 0$) were all repeated 10 times with different IVS threshold values. IVS_0 was varied from 0.90 to 0.99 in steps of 0.01. For each condition, the run with the smallest variance was taken. Only runs that led to a successful direction estimation of all speakers and where at least 2% of the samples passed the IVS filter mask were considered.

For two speakers, a small variance was achieved for any integration time from 2.5 to 20 cycles. When a larger number of speakers was simultaneously active, temporal integration was more critical and the highest precision was found for $\tau_s = 5 \cdot T_c$. This integration time corresponds to 5 ms in the 1-kHz band and 10 ms in the 500-Hz band. The runs that resulted in the smallest variance at $\tau_s = 5 \cdot T_c$ had IVS_0 = 0.98 for 4-5 speakers and IVS_0 = 0.99 for 2-3 speakers. IVS_0 = 0.99 did not lead to successful localization for more than three speakers. Therefore the parameters for all following experiments were frozen at $\tau_s = 5 \cdot T_c$ and IVS_0 = 0.98.

The evaluation of the histogram variance for the speaker at -30° led to similar results. The variance was always smaller by a factor of 2-3 which was sometimes only the width of one bin. This made the Gaussian fit procedure difficult.

The precision of the high-frequency modulation IPDs was worse than the fine-structure precision in all conditions. Modulation IPDs were therefore not considered in this evaluation and were investigated separately Sec. 5.4.2.

5.4.2 Influence of modulation IPDs

Due to the identical model structure for the fine-structure and the modulation path, all IVS parameters ($\tau_s = 5 \cdot T_c$, IVS$_0$ = 0.98) were kept unchanged for the investigation of modulator IPDs. Here, the cycle duration T_c, refers to the center frequency of the modulation filter.

Figure 4 shows an example of two concurrent speakers. The width of the two azimuth distributions from modulation filters at 135 and 216 Hz (Fig. 5.4a and 5.4b, respectively) was quite broad. For the direction estimate, the width was, however, of less concern than in the case of the fine-structure based histograms since only a single direction has to be extracted from each histogram: The tuning of the modulation filters to the fundamental frequency of the speakers already pre-filters the data. Still, the performance of envelope based DOA estimates did not reach that of the fine-structure based estimates. It was not possible to estimate the direction of four or five speakers solely on envelope IPDs.

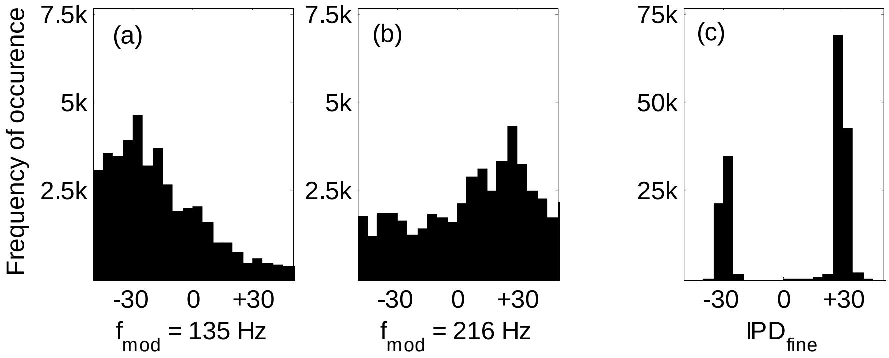

Fig. 5.4 - Pooled IVS-filtered DOA histograms for two concurrent speakers. The first speaker was located at α = -30° and had a median pitch frequency f_0 = 216 Hz. The second speaker was located at α = +30° had a median pitch frequency f_0 = 135 Hz. Panels (a) and (b) show the histograms based on the modulator IPD. The modulation filters were centered at 216 Hz (a) and at 135 Hz (b). Both histograms originate from the pooled estimates of 11 auditory filters in the range of 1.4 to 5.0 kHz. Panel (c) shows the pooled histogram resulting from the 12 low-frequency fine-structure filters.

5.4.3 Influence of noise and reverberation

So far, the model was tested under ideal conditions without any noise or reverberation. However, in daily-life situations of speech communication, listeners have to deal with these disturbing influences. As mentioned in the introduction, normal hearing listeners often outperform technical systems under such adverse conditions. The influence of omnidirectional, speech-shaped noise on the direction estimation was tested for one, two, and three non-moving speakers at two different signal-to-noise ratios (SNR). The SNR was defined as the ratio of the energy from the speaker at α = 30° to the total noise energy. For the two and three speaker condition, the second speaker from α = 0° and the third speaker from α = -30° were added at the same presentation level as the first speaker. Thus their energy was not considered in the SNR calculation.

Fig. 5.5 shows several histograms of the direction estimation both without (first row) and with (second row) the preceding IVS filtering. An estimation of the target direction was always possible at 0 dB SNR (left-hand block in Fig.5) with and without IVS mask. The histograms with IVS mask (lower row) show always more precise peaks. At -6 dB SNR (middle block), however, only a single speaker can be detected in the unfiltered conditions. The three tiny peaks in the unfiltered histogram for three speakers are not sufficient for three direction predictions. The right peak for instance is caused by the noise rather than a speaker. With IVS filtering, direction estimation is still possible, especially for several speakers. For a single speaker at -6 dB SNR (lower left panel of the middle block) there is a significant amount of noise passed through the IVS filter which leads to the only condition in this study where filtering decreases the robustness. For two and three speakers, however, the additional speakers mask the noise and the predictions become more precise when additional speakers are present. Erratic direction estimates would be also obvious in the IVS-filtered noise alone condition (lower panel of the right-hand block). Further analysis indicated that the performance was independent of the temporal integration time in the range of 5 to 20 cycle durations (not shown). The time constant was therefore kept unchanged from the first evaluation with $\tau_s = 5 \cdot T_c$.

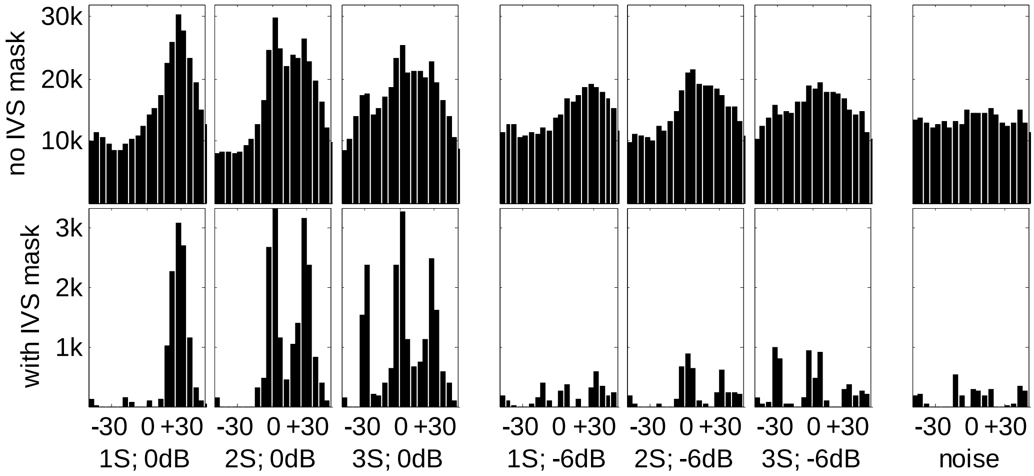

Fig. 5.5 - Pooled DOA histograms of different speaker-in-noise conditions. The noise was always omnidirectional and had a speech-shaped spectrum. The first speaker was located at +30°, the second at 0°, and the third at -30°. Histograms are drawn without IVS filtering (upper row) and with IVS filtering (lower row). Left block: 1, 2, and 3 speakers at 0 dB SNR. Middle block: same speakers at -6 dB SNR. Right column: omnidirectional noise alone.

A binaural room impulse response recorded in a small-sized office room (4 x 5.5 m, T_{60}~ 350 ms) was used in order to investigate the effect of reverberation on the DOA estimation in the model. A single speaker was positioned at +45° with respect to the coordinate system of the dummy head. Fig. 5.6a shows the pooled fine-structure IPD histogram with IVS filtering, resulting from the DOA estimation. It can be seen that the correct 45° estimate interferes with erroneous estimates from about 10 to 30°. These erroneous estimates could be assigned to originate from low-frequency room modes, since the dummy head was placed between two plain reflecting surfaces (window and wall with closed door). In the HRIRs, high-energy modes were found up to 280 Hz. To reduce the effect of the room modes, Fig. 5.6b shows the pooled histogram only for frequency channels above 500 Hz were no distinct modes were obvious in the BRIR. A considerably less disturbed direction estimation was observed in comparison to Fig. 5.6a. Panel (c) shows the histogram only based on the low-frequency channels below 400 Hz. Here the azimuth values were dominated by the IPD of the room mode. Modulator IPD-based DOA estimation is shown in Fig. 5.6d. Hardly any direction estimate was possible, which was an even worse performance than in the previous experiments without reverberation. The number of samples that passed the IVS filter was about the same for this office room condition with a single speaker as for five speakers in the free-field condition.

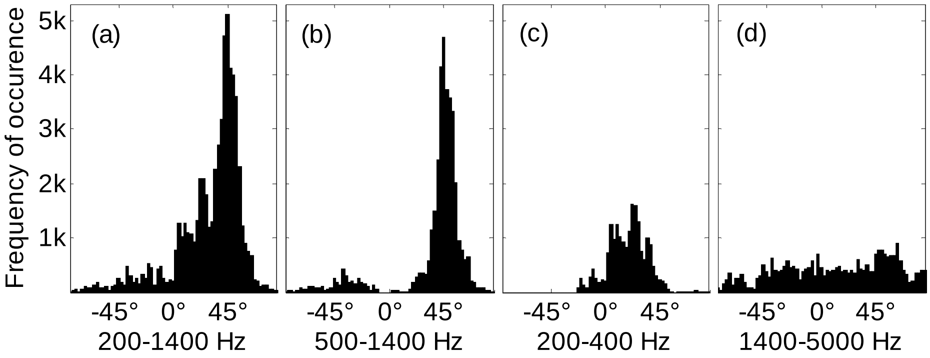

Fig. 5.6 - Pooled DOA histograms including IVS filtering for an office room condition with a single speaker at 45°. Panel (a) shows the pooled histogram for the full frequency range of 200-1400 Hz as used in the previous evaluation conditions. The middle panels show the pooled histograms with restricted frequency regions of 500-1400 Hz (b) and 200-400 Hz (c), respectively. The correct direction is well represented in the high-frequency region (a), while the estimates are dominated by the effect of room modes in the low frequency region (b). Panel (d) shows the pooled envelope-based histogram from the frequency region above 1.4 kHz. No direction estimate is could be derived from the histogram in this case.

Further tests (not shown) revealed that it was only possible to estimate directions of up to two concurrent speakers in this office room when the optimal frequency range of 500-1400 Hz was used. In the corresponding condition with two speakers in free field, the histogram peaks were on average 10 times as high as in this reverberant condition, indicating a very strong influence of even short reverberation times on the IVS.

5.4.4 Tracking of moving speakers

In real-life situations the interaural parameters of the target source are usually not constant. Movement of the listener, rotation of the listeners head or movement of the target source lead to temporally varying interaural parameters. The time scales of these changes are usually in the order of hundreds of milliseconds to seconds for movement of the listener or sources and considerably smaller for head rotations of the listener. Overall these are slower than the rapid changes in the interaural parameters caused by concurrent speakers where the IPD jumps from values caused by one speaker to values caused by another or by perturbations from ambient noise and reverberation. So far the direction information was always collected for stationary speakers for the duration of a complete sentence. This section now investigates the ability to follow moving speakers.

It is assumed that the usually short time segments in which the IVS exceeds IVS_0 are recognized as one localization event or “glimpse”. This assumption is based on a similar hypothesis from speech recognition called “glimpsing speech” (Cooke 2003). The azimuth estimations were therefore averaged over the total duration of a glimpse - usually 1 to 20 ms. In the rare occasion that the above threshold duration was longer than 50 ms, a new glimpse is assigned every 50 ms. If several glimpses are detected from the same direction in a reasonably short interval, the probability that these glimpses really indicate a source rather than noise or erroneous estimates is very high. For instance the probability that eight or more out of 100 glimpses are localized in the same 10° interval by chance is less than 1 % in all tested conditions. This was derived from a multinomial distribution with 18 bins. The same holds for 6 glimpses out of 50. When six glimpses were detected for all speaker directions, the total number of glimpses was usually still below 50. For eight glimpses the total number was always below 100. Therefore six to eight glimpses establish a very secure indicator that there is really a source from the respective direction. In Fig. 5.7a it can be seen that the necessary time to collect this number of glimpses is about 0.25 s for two speakers, 0.75 s for three speakers and even 1.25 s for two speakers in noise with 0 dB SNR.

In the second part of this section an example of a speaker tracking is shown. The task of a source tracking mechanism is to group glimpses to their corresponding sources and track them in a dynamic fault-tolerant model. The lower limit for the required grouping time follows from the first part of this section. If the source movement within this grouping time is lower than the distance between concurrent sources a tracking of sources is in theory possible.

This example was realized by utilizing an existing framework toolbox of so-called Rao-Blackwellized particle filters (Särkkä et al. 2007). Three moving speakers were simulated in the virtual acoustic space. The first speaker moved from the right to the center, the second speaker from the center to the left, and the third speaker from the left to the right. Particles were initialized in the range from -90° to +90°. The best model results were achieved with the parameters sigma = 3° and F(1,2) = 4 (see Hartikainen and Särkkä 2008 for details). All other parameters were kept unchanged at default. The output of the tracking is shown in Fig. 5.7b. With the given parameters, the creation of new speaker estimates was rather conservative. This results in a relatively long time until all speakers were detected though it had the advantage that phantom sources hardly ever occur.

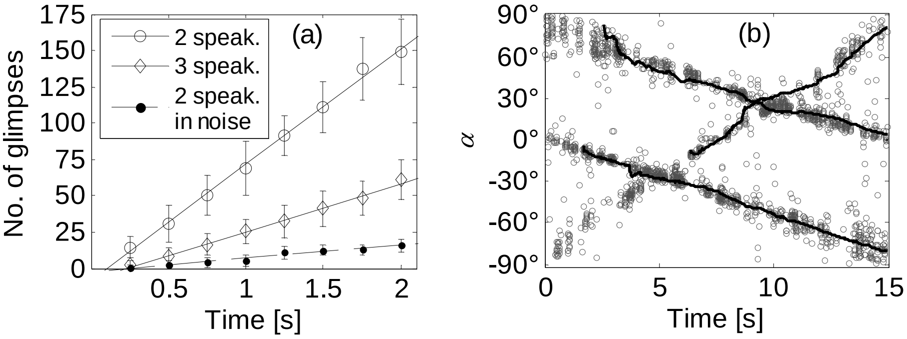

Fig. 5.7 - Panel (a): Number of DOA estimate glimpses collected for varying observation times. The values indicate the minimum of collected glimpses for two or three speakers in silence and two speakers in noise. Glimpses were collected in the ± 5° range around the true speaker directions. The lines indicate a linear fit to the data. Panel (b): The grey circles indicate DOA glimpses for a three moving speaker configuration. The solid lines are the tracking estimates of three moving speakers with a particle filter toolbox of Särkkä et al. (2007).

5.5 Discussion

A model for direction of arrival (DOA) estimation was suggested considering the four aspects of auditory processing, i.e., (i) temporal resolution, (ii) limited fine-structure information at high frequencies, (iii) temporal envelope disparities, and (iv) the limited ITD range. The implications of each aspect on model architecture and model performance are discussed in the following.

The temporal resolution was realized by a low-pass filtering the interaural transfer function and the interaural vector strength (IVS) as a substitute of the interaural coherence (IC). Interaural coherence as a measure to identify time-frequency instants corresponding to a distinct, located sound source, was introduced by Allen et al. (1977) and implemented for an auditory model by Faller and Merimaa (2004). In line with the study of Faller and Merimaa (2004) a big increase in estimation precision was achieved with an IVS or IC filter mask, especially in the presence of noise or several concurrent speakers. The integration time constant for extraction of the IC in Faller and Merimaa (2004) was fixed at 10 ms based on Lindemann (1986). Faller and Merimaa stated that there was further need for a frequency dependent IC threshold, since the short integration time is not suited to estimate the coherence in low-frequency bands where the cycle duration is close to the integration time. In the current study, the integration time constant for calculating the IVS was set to multiples of the cycle duration of the respective filter, allowing a frequency independent threshold. The use of a frequency dependent integration time constant and a frequency independent threshold is in line with the psychoacoustic finding that sensitivity to binaural decorrelation (e.g. detection of dichotic tones in noise) does not depend on the center frequency (e.g., van de Par and Kohlrausch 1999), while binaural temporal resolution depends on the center frequency in the fine-structure domain (Dietz et al., 2008).

It was further analyzed which temporal integration resulted in the best precision of the direction estimates in a multiple speaker condition. Independent of the speaker number, an integration time of five cycle durations led to the best precision. For speech in noise no critical dependence on the smoothing time constant was found. Five cycle durations translate to 10 ms in the 500-Hz band and 5 ms in the 1-kHz band. On average, these durations are just a little below most of the time constants in previously published localization models (e.g., Faller and Merimaa, 2004; Roman and Wang, 2008) and considerably lower than earlier estimates of binaural temporal resolution (30-200 ms, Culling and Colburn 2000; Kollmeier and Gilkey 1990) as used in binaural models (e.g., 30 ms, Breebaart et al., 2001a). The finding that a very fast feature extraction is required, particularly for more than two concurrent speakers, is related to the very short durations in which a single speaker dominates over all others. This result supports the assumption that the primary binaural temporal resolution is very high (Siveke et al., 2008; Dietz et al., 2008) and smoothing of interaural parameters or sluggishness occurs dependent on the specific task performed by the listener.

Another much longer time scale is necessary to estimate all directions of several concurrent speakers. For three speakers this time was estimated to be about 0.75 s. This is in line with psychoacoustic studies by Drullman and Bronkhorst (2000) who found an increase in reaction time of about 1.4 s until humans estimated the direction of a target speaker when two interfering talkers were added. Drullman and Bronkhorst also argue that the increase is due to the time until glimpses occur in which the binaural information can be extracted, since the increase in a monaural detection experiment is only 0.4 s. It can be argued that the difference of monaural and binaural increase is the “pure binaural effect” and this 1.0 s is only slightly longer than the model needed to get 99% probability.

The second aspect to be discussed is the inability of human listeners to exploit binaural information from fine-structure disparities in auditory frequency bands above about 1.4

or 1.5 kHz (e.g., Kuhn, 1977). The typical low-pass filtering for modeling the hair cell transduction process, followed by an internal additive noise source at hearing threshold does only damp the fine-structure information but does not guarantee that the binaural model processor cannot still exploit it (e.g., Breebaart et al., 2001a; Faller and Merimaa, 2004; Dietz et al., 2008). In this case it may even happen that the presentation level has a strong influence on the model performance, because detection accuracy is only limited by the noise at hearing threshold. Since the current model separates fine-structure and envelope IPDs anyway, it was possible to simply discard all fine-structure IPDs in frequency channels above 1.4 kHz. In the free field, the model performs well even with this limited frequency range. However, in reverberant environments technical models (e.g., Rohdenburg et al., 2008) and auditory models without fine-structure restriction (e.g., Roman and Wang, 2008) seem to be more robust. Beside ongoing binaural cues from IPDs and ILDs, it has been shown that onset cues are very important for both human and model performance (Braasch 2002; Supper et al. 2006), especially in reverberant environments. The current model implementation gives more weight to onset cues as a consequence of the IVS mask; however, the total weighting of onsets cannot be controlled and is probably rather small in a single continuously spoken sentence. Another helpful mechanism of the auditory system is the suppression of early reflections (precedence effect). Such a mechanism was not implemented explicitly in this model or in any of the previously cited models, but this can partially be accounted for by the integration window and filter ringing. Other mechanisms of extracting the ITD, such as zero-crossing statistics (e.g., Park and Stern 2009) may also be more robust for speech stimuli. Choosing the IPD as primary binaural feature was motivated by the limited ITD range of the mammalian auditory system, not by purely technical considerations (see fourth aspect below).

Even so it was just argued that fine-structure IPDs can not be exploited at higher frequencies, it is known that temporal disparities in the stimulus envelope do mediate spatial hearing (e.g., Bernstein and Trahiotis, 1994). This third aspect was accounted for by deriving azimuth estimates from modulator IPDs in the frequency region of 1.4 to 5.0 kHz. An envelope-frequency selective process was employed by filtering the pitch region of the target speaker from the modulation region. A band-pass filter centered at the average fundamental frequency of the target speaker was used to extract the envelope-based IPD. This process “logs” specifically on the fundamental frequency of a target speaker which determines the main envelope fluctuation rate during voiced parts of the speech. If interfering speakers with a different fundamental frequency are present, the signal-to-interferer ratio is increased and the interaural disparities mostly represent the direction of the target speaker. Due to the long cycle durations of the modulation frequencies it was not possible to estimate several speakers in the same modulation filter. Therefore the a-priori pitch information was important for the modulator IPD based estimation. For future applications without a-priori pitch information, it would be desirable to combine the model with a multipitch tracking algorithm (e.g., Wu et al. 2003).

Overall the estimation accuracy based solely on modulation IPDs is still worse than the performance with fine-structure IPDs. This is in line with psychoacoustic experiments of the just noticeable interaural time difference (JND). While JNDs for modulator ITDs are usually equal or above 100 µs (e.g., Bernstein and Trahiotis 2002), fine-structure ITD based JNDs are typically below 50 µs and can drop to about 10 µs for trained listeners (e.g., Koehnke et al. 1986). For sources from straight ahead the ITD changes about 10 µs per degree azimuth, leaving a 10° azimuth JND for undisturbed optimal modulator disparities. The model was tested under worse conditions with a direction error not exceeding 10°. A higher precision is difficult since modulation frequencies can only be exploited up to about 250 Hz (e.g., Bernstein and Trahiotis, 2002). At these low frequencies a 100-µs ITD translates into an IPD of π/20 or less and JNDs are large even for pure tones. The difference in accuracy for fine-structure and envelope-based direction estimation appears therefore to be directly related to the different maximum frequencies that can be used: 1.4 kHz for the fine-structure and about 250 Hz for the modulator IPD. In the model, the modulation frequencies have longer integration times of 20 to 50 ms ($\tau_s = 5 \cdot T_c$). Shorter integration times would disturb the IVS mask, longer integration times would again reduce the capability to capture the short but important segments in which one speaker dominates over several others. In all of the tested conditions, the direction estimate from the modulation filters could not improve the overall performance when combined with the fine-structure based direction estimate.

The fourth aspect is the limited internal delay range found in electrophysiological recordings in mammals. The inter delays seem to be restricted to the half of a cycle duration at neurons best frequency (e.g., McAlpine et al. 2001). For a 1-kHz neuron this would only be 500 µs. The common cross-correlation models with at least 750-µs maximal time-delay, would therefore extract information which can probably not be assessed by the human auditory system. Psychoacoustic studies (e.g., Sayers, 1964) indicate that pure tones with large IPDs offer indeed two possible direction estimates at $\alpha_1 = p_f\left(|\mathrm{IPD}|\right)$ and at $\alpha_2 = p_f\left(|2\pi - \mathrm{IPD}|\right)$. The ILD based selection of the two possibilities increased the range of unambiguous IPD-based DOA estimation from 700 to 1400 Hz. Now both phase-locking and the internal delay range limit fine-structure based localization to the same frequency region, and no further information is discarded by considering the π-limit.

5.6 Conclusions

The focus was put on four auditory aspects (i) temporal resolution, (ii) limited phase-locking, (iii) temporal envelope disparities, and (iv) the limited ITD range. A binaural localization model was built around these aspects and their influence on localization performance led to the following conclusions:

1. A high temporal resolution is necessary to localize several concurrent speakers. An integration time constant of five cycles of the respective center frequency of each filter band yields the best results and allows for a frequency independent interaural coherence filter mask. In free field five concurrent speakers can be localized with maximum errors of less than 5°.

2. Two concurrent speakers can be localized solely based on information from high-frequency envelope IPDs. Modulation filters tuned to the fundamental frequency of the target speaker suppress the influence of the other speaker almost completely. The precision and the robustness against noise or reverberation is worse than for fine-structure based direction estimates.

3. The sign of the ILD can be employed to unwrap the IPD to an unambiguous range from [-2π; 2π]. With this larger range, fine-structure based azimuth mapping is possible up to about 1.4 kHz, even if internal delays are π-limited. The 1.4 kHz limit roughly coincides with the human ability to exploit temporal fine-structure disparities and does therefore not result in any additional limitation.

Chapter 6

Summary and concluding remarks

The IPD model introduced in chapter 2 has proven to be suitable to account for a large variety of psychoacoustic data from both own experiments and literature. The most important results from the own experiments are:

- The temporal resolution for binaural beat detection was found to be mainly limited by the bandwidth of the auditory filters (chapter 3). In line with very recent findings by e.g., Siveke et al. (2008) this was much faster than traditionally assumed.

- The lateralization caused by a pure modulator IPD is independent of the modulation frequency in the tested range of 25 to 100 Hz (chapter 4). This is equivalent to a factor 4 difference in lateralization in terms of modulator ITD between the 25 Hz and the 100 Hz modulation frequency.

- When carrier and modulator IPDs of sinusoidally amplitude modulated (SAM) tones were set in opposition, the largest modulator IPD required to compensate for a carrier IPD occurred when the carrier IPD was 45°. This was in strong contrast to matching experiments, where the strongest lateralization was usually found at 135° (chapter 4).

- Strong subject- and level-dependences were found for matching the lateralization of time delayed SAM tones (chapter 4).

A model solely based on a delay line with half-cycle restriction failed to account for relevant psychoacoustic data (Thompson et al., 2006). The problem is that temporal envelope disparities influence psychoacoustic results but they cannot be captured by a single π-limited channel. In that respect the current IPD model can be seen as an extension by a second temporal channel extracting the demodulated envelope frequencies directly from the half-wave rectification of the hair cell transduction. With this additional channel it was possible to account for binaural release from masking for transposed tones in chapter 3 and to diminish the problem Thompson et al. (2006) had pointed out (chapter 4). Each channel of the IPD model is half-cycle or π-restricted by definition and the model can therefore be seen as a pioneering effective binaural model based on mammalian physiology and human psychoacoustics.

The main difference between direct detection of interaural modulation disparities and the conventional modulation sensitivity via interaural group delay (Stern et al., 1988) is the bandwidth limitation introduced by the width of the auditory filters. While demodulated envelope disparities are within channel cues, detection of interaural group delay is an across channel process. Future experiments and modeling of bandwidth-dependent lateralization may reveal further insights on the true processing of temporal envelope disparities. In chapter 4, experiment 4 and 5 revealed some weakness of the IPD model to account for dependence of lateralization on envelope shape and overall level for all subjects. These results have triggered an additional study on “the role of envelope waveform, adaptation, and attacks in binaural perception” (Ewert et al., in press). The experiments revealed a different importance of rise and decay flanks and a strong influence of short pause durations prior to the rising flanks. Both observations could be accounted for by fast adaptation on a time-scale of about 5 ms. Including an adaptive weighting of instantaneous differences in the modulation branch of the IPD model should be investigated in the ongoing work. It may allow a completely fused processing of envelope ITDs and ILDs based on excitatory-inhibitory processes as it would be even more realistic for a physiologic model of lateral superior olive processing: Temporal differences lead to temporally fluctuating instantaneous level differences but without adaptive weighting these differences cancel out over time. With adaptive weighting the differences do not cancel out - like a static interaural level difference.

The second novelty of the IPD model is that interaural timing disparities are derived without any time delay compensation. At the time the model was developed there was no physiological data available answering whether there is really no time delay compensation and just a phase comparison or if there is a limited and frequency dependent characteristic time delay compensation. The answer of this question is not of critical importance for the IPD model, because even if characteristic time delays exist up to the half of a cycle duration, they could not transmit enough information about the stimulus envelope to account for psychoacoustic data (Thomson et al., 2006). Nevertheless it is worth mentioning that after the IPD model was developed a study by Agapiou and McAlpine (2008) was published, showing that it is frequency dependent time delay and not a characteristic phase. Therefore the assumption in chapter 4 that the frequency independent best IPD indicates a pure phase and no time delay compensation is too strict and the model could additionally be allowed to compensate for 45° delays. Another interesting result from Agapiou and McAlpine (2008) is that any additional shift of some tuning curves from 45° towards 180° is employed by a characteristic phase response and not by additional time delay compensation. This result is another indication against a chain with different time delays.

Besides modeling psychoacoustic data of normal-hearing subjects the model may also be suitable to be extended (or downsized) for data from subjects with synchronized bilateral cochlea implants (Majdak et al., 2006). Caused by the specific stimulus design in cochlea implants with discreet pulses as substitute for the carrier, the notion of “carrier” IPD and modulator IPD is also obvious in electric hearing. Furthermore the absence of a cochlea allows a very direct control of the pre-processing but it also disables the phase-locking to temporal fine-structure.

The possibilities of the model for direction of arrival (DOA) estimation have been demonstrated in chapter 5. Psychoacoustic findings from chapter 3 about the high temporal resolution and established auditory side conditions such as the limitation of exploiting the fine-structure at high frequencies led to very different strategies for the DOA estimation. While generally good and competitive, the model did not reach the standard of technical DOA estimators in the presence of reverberation. Engineers have put a lot of effort in smart integration of information over frequencies since quite some time (e.g., Knapp and Charter, 1976; Roman and Wang, 2008). On the other hand the auditory system probably performs the frequency integration almost optimally for each specific task, which can include very different strategies. This task dependence makes it hard to model it in an auditory manner but since the processing is probably done in the auditory cortex it is not in the focus of a model for the binaural pathway at the level of the brain stem or mid brain. Nevertheless, it may be interesting to test integration techniques from engineering models if there is the need for a quantitative performance comparison to humans or technical models. More in the focus of future model-based DOA estimators is probably an explanation for the precedence effect and the salience of interaural time difference at the stimulus onsets. The detailed but mostly descriptive way in which these cues have been studied in the past did not result in a model or in a precise understanding of the underlying principles. A model-based approach of central sub-cortical processing possibly reveals new feature extraction schemes, as did the DOA model with ongoing interaural parameters in chapter 5.

In all versions the modeling was performed in feed-forward manner. This is by no means an argument against feedback or efferent processes. Of course efferent processes exist and it is very likely that they play an important role in processes, which have been simplified to feed-forward in this model. However, from the perspective of an effective model it is desirable to describe the data with as little free parameters and as simple as possible. Even with the additional objective of this model not to contradict mammalian physiologic data it was not intended to include any circuit without having a clear idea about its benefits. In the related study of Ewert et al. (in press) the most successful implementation was also a feed-forward adaptation.

Altogether the work combined several branches of binaural research: modeling, psychoacoustic measurements, integration of physiologic data, and DOA estimation as application. The same model groundwork could successfully be employed for all different tasks. The model offers an intuitive processing of familiar cues from modern psychoacoustics such as fine-structure and envelope disparities (e.g., Eddins and Barber, 1998; Smith et al., 2002; Furukawa, 2008). The cited studies would not fit with existing one-channel models but they are calling for a two-channel model. The IPD model may therefore help to bring new live into the mutual support of psychoacoustics and modeling. On the other hand it calls for neurocomputation to model the binaural interaction with more physiologic accuracy but again in mutual support with effective modeling and psychoacoustics. An accurate model of the human binaural system may help to identify the stages and parameters that are affected by different kinds of hearing losses, similar to a recent monaural model by Jepsen et al. (2008). In the same way effective binaural modeling is important to hold (or bring) binaural psychoacoustic and physiologic research closer together and therewith to also strengthen research in applications such as binaural hearing aids or cochlea implants. With the development and application of the IPD model this thesis intended to be useful for further applied and fundamental binaural research.

Bibliography

Agapiou, J. P., and McAlpine, D. (**2008**), ‘Low-frequency envelope sensitivity produces asymmetric binaural tuning curves,’ *J Neurophysiol* **100**, 2381–2396.

Ajmera, J., Lathoud, G. and McCowan, L. (**2004**), Clustering and segmenting speakers and their locations in meetings, *in* ‘Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004),’ **1**, 605–608.

Allen, J. B., Berkley, D. A. and Blauert, J. (**1977**), ‘Multimicrophone signal-processing technique to remove room reverberation from speech signals,’ *J Acoust Soc Am* **62**(4), 912–915.

Batra, R., Kuwada, S. and Fitzpatrick, D. C. (**1997**), ‘Sensitivity to interaural temporal disparities of low- and high-frequency neurons in the superior olivary complex. I. Heterogeneity of responses,’ *J Neurophysiol* **78**(3), 1222–1236.

Batra, R. and Yin, T. C. T. (**2004**), ‘Cross correlation by neurons of the medial superior olive: a reexamination,’ *J Assoc Res Otolaryngol* **5**(3), 238–252.

Bernstein, L. R. and Trahiotis, C. (**1982**), ‘Detection of interaural delay in high-frequency noise,’ *J Acoust Soc Am* **71**(1), 147–152.

Bernstein, L. R. and Trahiotis, C. (**1985*a***), ‘Lateralization of low-frequency, complex waveforms: the use of envelope-based temporal disparities,’ *J Acoust Soc Am* **77**(5), 1868–1880.

Bernstein, L. R. and Trahiotis, C. (**1985*b***), ‘Lateralization of sinusoidally amplitude-modulated tones: effects of spectral locus and temporal variation,’ *J Acoust Soc Am* **78**(2), 514–523.

Bernstein, L. R. and Trahiotis, C. (**1994**), ‘Detection of interaural delay in high-frequency sinusoidally amplitude-modulated tones, two-tone complexes, and bands of noise,’ *J Acoust Soc Am* **95**(6), 3561–3567.

Bernstein, L. R. and Trahiotis, C. (**1996**), ‘The normalized correlation: accounting for binaural detection across center frequency,’ *J Acoust Soc Am* **100**(6), 3774–3784.

Bernstein, L. R. and Trahiotis, C. (**2002**), ‘Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli",’ *J Acoust Soc Am* **112**(3), 1026–1036.

Bernstein, L. R. and Trahiotis, C. (**2007**), ‘Why do transposed stimuli enhance binaural processing?: Interaural envelope correlation vs envelope normalized fourth moment,’ *J Acoust Soc Am* **121**(1), EL23–EL28.

Beutelmann, R. and Brand, T. (**2006**), ‘Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners,’ *J Acoust Soc Am* **120**(1), 331–342.

Blauert, J. and Lindemann, W. (**1986**), ‘Auditory spaciousness: Some further psychoacoustic analyses,’ *J Acoust Soc Am* **80**(2), 533–542.

Blodgett, H. C., Wilbanks, W. A. and Jeffress, L. A. (**1956**), ‘Effect of large interaural time differences upon the judgment of sidedness,’ *J Acoust Soc Am* **28**(4), 639–643.

Boehnke, S. E., Hall, S. E. and Marquardt, T. (**2002**), ‘Detection of static and dynamic changes in interaural correlation,’ *J Acoust Soc Am* **112**(4), 1617–1626.

Borisyuk, A., Semple, M. N. and Rinzel, J. (**2002**), ‘Adaptation and Inhibition Underlie Responses to Time-Varying Interaural Phase Cues in a Model of Inferior Colliculus Neurons,’ *J Neurophysiol* **88**, 2134–2146.

Braasch, J. (**2002**), ‘Localization in the Presence of a Distracter and Reverberation in the Frontal Horizontal Plane. II. Model Algorithms,’ *Acta acustica / Acustica* **88**, 956–969.

Brand, A., Behrend, O., Marquardt, T., McAlpine, D. and Grothe, B. (**2002**), ‘Precise inhibition is essential for microsecond interaural time difference coding,’ *Nature* **417**, 543–547.

Breebaart, J., van de Par, S. and Kohlrausch, A. (**2001*a***), ‘Binaural processing model based on contralateral inhibition. I. Model structure,’ *J Acoust Soc Am* **110**(2), 1074–1088.

Breebaart, J., van de Par, S. and Kohlrausch, A. (**2001*b***), ‘Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters,’ *J Acoust Soc Am* **110**(2), 1089–1104.

Breebaart, J., van de Par, S. and Kohlrausch, A. (**2001*c***), ‘Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters,’ *J Acoust Soc Am* **110**(2), 1105–1117.

Buell, T. N., Griffin, S. J. and Bernstein, L. R. (**2008**), ‘Listeners’ sensitivity to "onset/offset" and "ongoing" interaural delays in high-frequency, sinusoidally amplitude-modulated tones,’ *J Acoust Soc Am* **123**(1), 279–294.

Calmes, L., Lakemeyer, G. and Wagner, H. (**2007**), ‘Azimuthal sound localization using coincidence of timing across frequency on a robotic platform,’ *J Acoust Soc Am* **121**(4), 2034–2048.

Carr, C. E. and Konishi, M. (**1988**), ‘Axonal delay lines for time measurement in the owl’s brainstem,’ *Proc Natl Acad Sci USA* **85**(21), 8311–8315.

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

Cherry, C. E. (**1953**), ‘Some experiments on the recognition of speech, with one and with two ears,’ *J Acoust Soc Am* **25**(5), 975–979.

Colburn, H. S. (**1977**), ‘Theory of binaural interaction based on auditorynerve data. II. Detection of tones in noise,’ *J Acoust Soc Am* **61**, 525–533.

Colburn, H. S. and Durlach, N. I. (**1978**), ‘Models of binaural interaction,’ in *Handbook of Perception*, edited by Carterette, E. C. and Friedman, M. P. (Academic Press, New York), pp. 467–518.

Cooke, M. (**2003**), ‘Glimpsing speech,’ *Journal of Phonetics* **31**, 579 – 584.

Cooke, M. P. and Brown, G. J. (**1993**), ‘Computational auditory scene analysis: exploiting principles of perceived continuity,’ *Speech Communication* **13**, 391–399.

Crow, G., Rupert, A. L. and Moushegian, G. (**1978**), ‘Phase locking in monaural and binaural medullary neurons: Implications for binaural phenomena,’ *J Acoust Soc Am* **64**(2), 493–501.

Culling, J. F. (**2007**), ‘Evidence specifically favoring the equalization-cancellation theory of binaural unmasking,’ *J Acoust Soc Am* **122**(5), 2803–2813.

Culling, J. F. and Colburn, H. S. (**2000**), ‘Binaural sluggishness in the perception of tone sequences and speech in noise,’ *J Acoust Soc Am* **107**(1), 517–527.

Culling, J. F. and Summerfield, Q. (**1998**), ‘Measurements of the binaural temporal window using a detection task,’ *J Acoust Soc Am* **103**(6), 3540–3553.

Dau, T., Kollmeier, B. and Kohlrausch, A. (**1997**), ‘Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers,’ *J Acoust Soc Am* **102**(5), 2892–2905.

Dau, T., Püschel, D. and Kohlrausch, A. (**1996**), ‘A quantitative model of the "effective” signal processing in the auditory system. I. Model structure,’ *J Acoust Soc Am* **99**(6), 3615–3622.

de Cheveigné, A. and Kawahara, H. (**1999**), ‘Multiple period estimation and pitch perception model,’ *Speech Communication* **27**, 175–185.

de Cheveigné, A. and Kawahara, H. (**2002**), ‘YIN, a fundamental frequency estimator for speech and music,’ *J Acoust Soc Am* **111**(4), 1917–1930.

Dietz, M., Ewert, S. D. and Hohmann, V. (**2009**), ‘Lateralization of stimuli with independent fine-structure and envelope-based temporal disparities,’ *J Acoust Soc Am* **125**(3), 1622–1635.

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 Res* **1220**, 234–245.

Dreyer, A. and Delgutte, B. (**2006**), ‘Phase locking of auditory-nerve fibers to the envelopes of high-frequency sounds: implications for sound localization,’ *J Neurophysiol* **96**(5), 2327–2341.

Drullman, R. and Bronkhorst, A. W. (**2000**), ‘Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation,’ *J Acoust Soc Am* **107**(4), 2224–2235.

Durlach, N. I. (**1963**), ‘Equalization and cancellation theory of binaural masking-level differences,’ *J Acoust Soc Am* **35**(8), 1206–1218.

Eddins, D. A. and Barber, L. E. (**1998**), ‘The influence of stimulus envelope and fine structure on the binaural masking level difference,’ *J Acoust Soc Am* **103**(5), 2578–2589.

Ewert, S. D. and Dau, T. (**2000**), ‘Characterizing frequency selectivity for envelope fluctuations,’ *J Acoust Soc Am* **108**(3), 1181–1196.

Ewert, S. D., Dietz, M., Klein-Hennig, M. and Hohmann, V. (in press), ‘The role of envelope wave form, adaptation, and attacks in binaural perception,’ in *Advances in auditory research: physiology, psychophysics and models*, edited by Lopez-Poveda, E. A., Palmer, A. R. and Meddis, R. (Springer, New York).

Faller, C. and Merimaa, J. (**2004**), ‘Source localization in complex listening situations: selection of binaural cues based on interaural coherence,’ *J Acoust Soc Am* **116**(5), 3075–3089.

Fitzpatrick, D. C., Batra, R., Stanford, T. R. and Kuwada, S. (**1997**), ‘A neuronal population code for sound localization,’ *Nature* **388**, 871–874.

Fletcher, H. (**1940**), ‘Auditory patterns,’ *Rev Mod Phys* **12**, 47–65.

Furukawa, S. (**2008**), ‘Detection of combined changes in interaural time and intensity differences: Segregated mechanisms in cue type and in operating frequency range?,’ *J Acoust Soc Am* **123**(3), 1602–1617.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (**1990**), ‘DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM,’ *U.S. Dept. of Commerce NTIS, Gaithersburg, MD* .

Goupell, M. J. and Hartmann, W. M. (**2006**), ‘Interaural fluctuations and the detection of interaural incoherence: bandwidth effects,’ *J Acoust Soc Am* **119**(6), 3971–3986.

Grantham, D. W. (**1982**), ‘Detectability of time-varying interaural correlation in narrow-band noise stimuli,’ *J Acoust Soc Am* **72**(4), 1178–1184.

Grimm, G., Hohmann, V. and Kollmeier, B. (**2009**), ‘Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-based noise reduction scheme,’ *IEEE Transactions on Audio, Speech, and Language Processing* **17**, 1408–1419.

Grothe, B., Covey, E. and Casseday, J. H. (**2001**), ‘Medial superior olive of the big brown bat: neuronal responses to pure tones, amplitude modulations, and pulse trains,’ *J Neurophysiol* **86**(5), 2219–2230.

Halverson, H. (**1922**), ‘Binaural localization of phase and intensity,’ *Am J Psychol* **33**, 178–212.

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

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

Hartikainen, J. and Särkkä, S. (**2008**), *Optimal filtering with Kalman filters and smoothers – a Manual for Matlab toolbox EKF/UKF*, Department of Biomedical Engineering and Computational Science, Helsinki University of Technology,. Version 1.2.

Haykin, S. and Chen, Z. (**2005**), ‘The cocktail party problem,’ *Neural Comput* **17**(9), 1875–1902.

Heil, P. (**2003**), ‘Coding of temporal onset envelope in the auditory system,’ *Speech Communication* **41**(1), 123–134.

Henning, G. B. (**1974**), ‘Detectability of interaural delay in high-frequency complex waveforms,’ *J Acoust Soc Am* **55**, 84–90.

Hess, W. (**2006**), Time-variant binaural-activity characteristics as indicator of auditory spatial attributes, PhD thesis, Bochum University, Germany.

Hirsh, I. (**1948**), ‘The influence of interaural phase on summation and inhibition,’ *J Acoust Soc Am* **20**, 536–544.

Hohmann, V. (**2002**), ‘Frequency analysis and synthesis using a Gammatone filterbank,’ *Acta acustica / Acustica* **88**, 433–442.

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

Jepsen, M. L., Ewert, S. D. and Dau, T. (**2008**), ‘A computational model of human auditory signal processing and perception,’ *J Acoust Soc Am* **124**(1), 422–438.

Joris, P. X. (**1996**), ‘Envelope coding in the lateral superior olive. II. Characteristic delays and comparison with responses in the medial superior olive,’ *J Neurophysiol* **76**, 2137–2156.

Joris, P. X. (**2003**), ‘Interaural time sensitivity dominated by cochlea-induced envelope patterns,’ *J Neurosci* **23**(15), 6345–6350.

Joris, P. X., van de Sande, B., Recio-Spinoso, A. and van der Heijden, M. (**2006**), ‘Auditory midbrain and nerve responses to sinusoidal variations in interaural correlation,’ *J Neurosci* **26**(1), 279–289.

Joris, P. X. and Yin, T. C. T. (**1995**), ‘Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences,’ *J Neurophysiol* **73**, 1043–1062.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V. and Kollmeier, B. (**2009**), ‘Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,’ *EURASIP J Adv Sig Proc* **2009**, 1–10. Article ID 298605.

Kiang, N. Y. (**1968**), ‘A survey of recent developments in the study of auditory physiology,’ *Ann Otol Rhinol Laryngol* **77**(4), 656–675.

Knapp, C. H. and Carter, G. C. (**1976**), ‘The generalized correlation method for estimation of time delay,’ *IEEE Transactions on Acoustics, Speech, and Signal Processing* **24**, 320–327.

Koehnke, J., Colburn, H. S. and Durlach, N. I. (**1986**), ‘Performance in several binaural-interaction experiments,’ *J Acoust Soc Am* **79**(5), 1558–1562.

Kohlrausch, A., Fassel, R. and Dau, T. (**2000**), ‘The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers,’ *J Acoust Soc Am* **108**(2), 723–734.

Kollmeier, B. and Gilkey, R. H. (**1990**), ‘Binaural forward and backward masking: Evidence for sluggishness in binaural detection,’ *J Acoust Soc Am* **87**(4), 1709–1719.

Kollmeier, B. and Koch, R. (**1994**), ‘Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,’ *J Acoust Soc Am* **95**(3), 1593–1602.

Kuhn, G. F. (**1977**), ‘Model for the interaural time differences in the azimuthal plane,’ *J Acoust Soc Am* **62**(1), 157–167.

Levitt, H. (**1971**), ‘Transformed up–down methods in psychoacoustics,’ *J Acoust Soc Am* **49**, 467–477.

Lindemann, W. (**1986**), ‘Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals,’ *J Acoust Soc Am* **80**(6), 1608–1622.

Liu, C., Wheeler, B. C., William D. O’Brien, J., Bilger, R. C., Lansing, C. R. and Feng, A. S. (**2000**), ‘Localization of multiple sound sources with two microphones,’ *J Acoust Soc Am* **108**(4), 1888–1905.

Liu, C., Wheeler, B. C., William D. O’Brien, J., Lansing, C. R., Bilger, R. C., Jones, D. L. and Feng, A. S. (**2001**), ‘A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers,’ *J Acoust Soc Am* **110**(6), 3218–3231.

Loftus, W. C., Bishop, D. C., Marie, R. L. S. and Oliver, D. L. (**2004**), ‘Organization of binaural excitatory and inhibitory inputs to the inferior colliculus from the superior olive,’ *J Comp Neurol* **472**(3), 330–344.

Louage, D. H. G., Joris, P. X. and van der Heijden, M. (**2006**), ‘Decorrelation sensitivity of auditory nerve and anteroventral cochlear nucleus fibers to broadband and narrowband noise,’ *J Neurosci* **26**(1), 96–108.

Majdak, P., Laback, B. and Baumgartner, W.-D. (**2006**), ‘Effects of interaural time differences in fine structure and envelope on lateral discrimination in electric hearing,’ *J Acoust Soc Am* **120**(4), 2190–2201.

Marquardt, T. and McAlpine, D. (**2007**), ‘A pi-limit for coding ITDs: Implications for binaural models,’ in *Hearing—From Sensory Processing to Perception*, edited by Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S. and Verhey, J. (Springer, Berlin), pp. 312–318.

McAlpine, D. and Grothe, B. (**2003**), ‘Sound localization and delay lines–do mammals fit the model?’ *Trends Neurosci* **26**(7), 347–350.

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

McFadden, D. and Pasanen, E. G. (**1976**), ‘Lateralization at high frequencies based on interaural time differences,’ *J Acoust Soc Am* **59**, 634–639.

Mehrgardt, S. and Mellert, V. (**1977**), ‘Transformation characteristics of the external human ear,’ *J Acoust Soc Am* **61**, 1567–1576.

Millman, R. E. and Bacon, S. P. (**2008**), ‘The influence of spread of excitation on the detection of amplitude modulation imposed on sinusoidal carriers at high levels,’ *J Acoust Soc Am* **123**(2), 1008–1016.

Mossop, J. E. and Culling, J. (**1998**), ‘Lateralization of large interaural delays,’ *J Acoust Soc Am* **104**, 1574–1579.

Nix, J. and Hohmann, V. (**2006**), ‘Sound source localization in real sound fields based on empirical statistics of interaural parameters,’ *J Acoust Soc Am* **119**(1), 463–479.

Nix, J. and Hohmann, V. (**2007**), ‘Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering,’ *IEEE Transactions on Audio, Speech, and Language Processing* **15**(3), 995–1008.

Oxenham, A. J. and Moore, B. C. J. (**1994**), ‘Modeling the additivity of nonsimultaneous masking,’ *Hear Res* **80**, 105–118.

Palmer, A. and Russell, I. (**1986**), ‘Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells,’ *Hear Res* **24**(1), 1–15.

Park, H.-M. and Stern, R. M. (**2009**), ‘Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero-crossings,’ *Speech Communication* **51**(1), 15–25.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J. and Rice, P. (**1987**), An efficient auditory filterbank based on the gammatone function., Paper presented at the Meeting of the IOC Speech Group on Auditory Modeling at RSRE, 14–15 December.

Peissig, J. and Kollmeier, B. (**1997**), ‘Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners,’ *J Acoust Soc Am* **101**(3), 1660–1670.

Phillips, D. P. (**2008**), ‘A perceptual architecture for sound lateralization in man,’ *Hear Res* **238**, 124–132.

Pollack, I. (**1978**), ‘Temporal switching between binaural information sources,’ *J Acoust Soc Am* **63**, 550–558.

Pollack, I. and Trittipoe, W. J. (**1959**), ‘Binaural listening and interaural noise cross correlation,’ *J Acoust Soc Am* **31**(9), 1250–1252.

Puria, S., Peake, W. T. and Rosowski, J. J. (**1997**), ‘Sound-pressure measurements in the cochlear vestibule of human-cadaver ears,’ *J Acoust Soc Am* **101**(5), 2754–2770.

Rayleigh, L. (**1907**), ‘On our perception of sound direction,’ *Phil Mag Series 6* **13**, 214–232.

Riedel, H. and Kollmeier, B. (**2006**), ‘Interaural delay-dependent changes in the binaural difference potential of the human auditory brain stem response,’ *Hear Res* **218**, 5–19.

Rohdenburg, T., Goetze, S., Hohmann, V., Kammeyer, K.-D. and Kollmeier, B. (**2008**), Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays, *in* ‘International Conference on Acoustics, Speech, and Signal Processing (ICASSP 08),’ Las Vegas, USA, pp. 2449–2452.

Roman, N. and Wang, D. (**2008**), ‘Binaural Tracking of Multiple Moving Sources,’ *IEEE Transactions on Audio, Speech, and Language Processing* **16**(4), 728–739.

Roman, N., Wang, D. and Brown, G. J. (**2003**), ‘Speech segregation based on sound localization,’ *J Acoust Soc Am* **114**, 2236–2252.

Ruggero, M. A. and Rich, N. C. (**1991**), ‘Furosemide alters organ of corti mechanics: evidence for feedback of outer hair cells upon the basilar membrane,’ *J. Neurosci.* **11**(4), 1057–1067.

Ruggero, M. A. and Rich, N. C. (**1997**), ‘Furosemide alters organ of Corti mechanics: evidence for feedback of outer hair cells upon the basilar membrane,’ *J Neurosci* **11**, 1057–1067.

Ruotolo, B. R., Stern, R. M. and Colburn, H. S. (**1979**), ‘Discrimination of symmetric time-intensity traded binaural stimuli,’ *J Acoust Soc Am* **66**(6), 1733–1737.

Sanchez-Longo, L. P., Forster, F. M. and Auth, T. L. (**1957**), ‘A clinical test for sound localization and its applications,’ *Neurology* **7**(9), 655–663.

Särkkä, S., Vehtari, A. and Lampinen, J. (**2007**), ‘Rao-Blackwellized particle filter for multiple target tracking,’ *Information Fusion* **8**(1), 2–15.

Sayers, B. M. (**1964**), ‘Acoustic-image lateralization judgments with binaural tones,’ *J Acoust Soc Am* **36**(5), 923–926.

Sayers, B. M. and Cherry, E. C. (**1957**), ‘Mechanism of binaural fusion in the hearing of speech,’ *J Acoust Soc Am* **29**(9), 973–987.

Siveke, I., Ewert, S. D., Grothe, B. and Wiegrebe, L. (**2008**), ‘Psychophysical and physiological evidence for fast binaural processing,’ *J Neurosci* **28**(9), 2043–2052.

Siveke, I., Ewert, S. D. and Wiegrebe, L. (**2007**), ‘Perceptual and physiological characteristics of binaural sluggishness,’ in *Hearing—From Sensory Processing to Perception*, edited by Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S. and Verhey, J. (Springer, Berlin), pp. 354–357.

Smith, Z. M., Delgutte, B. and Oxenham, A. J. (**2002**), ‘Chimaeric sounds reveal dichotomies in auditory perception,’ *Nature* **416**, 87–90.

Stern, R. M. and Colburn, H. S. (**1978**), ‘Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position,’ *J Acoust Soc Am* **64**(1), 127–140.

Stern, R. M. and Shear, G. D. (**1996**), ‘Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay,’ *J Acoust Soc Am* **100**(4), 2278–2288.

Stern, R. M., Zeiberg, A. S. and Trahiotis, C. (**1988**), ‘Lateralization of complex binaural stimuli: A weighted-image model,’ *J Acoust Soc Am* **84**(1), 156–165.

Supper, B., Brookes, T. and Rumsey, F. (**2006**), ‘An auditory onset detection algorithm for improved automatic source localization,’ *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 1008–1017.

Thompson, E. R. and Dau, T. (**2008**), ‘Binaural processing of modulated interaural level differences,’ *J Acoust Soc Am* **123**(2), 1017–1029.

Thompson, S. K., von Kriegstein, K., Deane-Pratt, A., Marquardt, T., Deichmann, R., Griffiths, T. D. and McAlpine, D. (**2006**), ‘Representation of interaural time delay in the human auditory midbrain,’ *Nat Neurosci* **9**(9), 1096–1098.

Thompson, S. P. (**1877**), ‘On binaural audition,’ *Phil Mag Series 5* **4**, 274–276.

Thompson, S. P. (**1882**), ‘On the function of the two ears in the perception of space,’ *Phil Mag Series 5* **13**, 406–416.

Trahiotis, C. and Bernstein, L. R. (**1986**), ‘Lateralization of bands of noise and sinusoidally amplitude-modulated tones: effects of spectral locus and bandwidth,’ *J Acoust Soc Am* **79**(6), 1950–1957.

Trahiotis, C. and Kappauf, W. E. (**1978**), ‘Regression interpretation of differences in time-intensity trading ratios obtained in studies of laterality using the method of adjustment,’ *J Acoust Soc Am* **64**, 1041–1047.

Trahiotis, C. and Stern, R. M. (**1989**), ‘Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase,’ *J Acoust Soc Am* **86**(4), 1285–1293.

van de Par, S. and Kohlrausch, A. (**1997**), ‘A new approach to comparing binaural masking level differences at low and high frequencies,’ *J Acoust Soc Am* **101**(3), 1671–1680.

van de Par, S. and Kohlrausch, A. (**1999**), ‘Dependence of binaural masking level differences on center frequency, masker bandwidth, and interaural parameters,’ *J Acoust Soc Am* **106**(1), 1940–1947.

Weiss, T. F. and Rose, C. (**1988**), ‘A comparison of synchronization filters in different auditory receptor organs,’ *Hear Res* **33**, 175–180.

Wittkop, T. and Hohmann, V. (**2003**), ‘Strategy-selective noise reduction for binaural digital hearing aids,’ *Speech Communication* **39**, 111–138.

Wu, M., Wang, D. and Brown, G. J. (**2003**), ‘A multipitch tracking algorithm for noisy speech,’ *IEEE Transactions on Audio, Speech, and Language Processing* **11**(3), 229–241.

Young, L. L. (**1976**), ‘Time-intensity trading functions for selected pure tones,’ *J Speech Hear Res* **19**, 55–67.

Young, L. L. and Carhart, R. (**1974**), ‘Time-intensity trading functions for pure tones and a high-frequency AM signal,’ *J Acoust Soc Am* **56**, 605–609.

Zerbs, C. (**2000**), Modeling the effective binaural signal processing in the auditory system, PhD thesis, Oldenburg University, Germany.

Danksagung

Ich danke allen Personen, die mich bei der Erstellung dieser Arbeit unterstützt haben und damit auch wesentlich zu meiner persönlichen Weiterentwicklung während dieser Zeit beigetragen haben.

Im Besonderen danke ich meinen drei Betreuern: Prof. Dr. Dr. Birger Kollmeier, PD Dr. Volker Hohmann und Dr. Stephan Ewert von deren wissenschaftlicher Brillanz ich in meinen ersten Jahren in der Forschung enorm profitiert habe. Sie haben mit ihrer hervorragenden und netten Betreuung und ihrem beeindruckenden Fachwissen meine Begeisterung für die Hörforschung geweckt.

Prof. Dr. Dr. Birger Kollmeier danke ich außerdem für die herzliche Aufnahme in die Arbeitsgruppe und dem damit verbundenen exzellenten Umfeld für Hörforschung.

Bei PD Dr. Volker Hohmann möchte ich mich zusätzlich für die interessante Fragestellung dieser Arbeit bedanken. Sowohl in allen fachlichen Belangen als auch für die Planung von Journal- und Tagungsbeiträgen war er ein hervorragender Mentor. Danke auch für die immer schnelle und unkomplizierte Hilfe und die ausgesprochen motivierende und freundschaftliche Atmosphäre in den vielen Diskussionen.

Bei Dr. Stephan Ewert möchte ich mich herzlich für die viele Zeit bedanken, die er während meines ersten Jahres mit mir vor dem Computer verbracht hat. Dabei habe ich sehr von seinem umfassenden Wissen über Signalverarbeitung profitiert. Außerdem hat sein unermüdlicher Einsatz beim Korrekturlesen der Veröffentlichungen wesentlich zu deren Erfolg beigetragen.

Bei Prof. Dr. David McAlpine möchte ich mich vielmals für die spontane Bereitschaft bedanken, sich als Prüfer für meine Disputation bereit zu erklären und die lange Anreise dafür auf sich zu nehmen. Außerdem danke ich ihm für die herzliche Gastfreundschaft bei meinem Besuch an der UCL sowie für viele motivierende Diskussionen.

Besonders bedanke ich mich außerdem bei Martin Klein-Hennig für die gute Zusammenarbeit, bei Dr. Helmut Riedel, Helge Lüddemann, Giso Grimm, Marc Nitschmann, Dr. Rainer Beutelmann und Prof. Dr. Jesko Verhey für inspirierende

binaurale Diskussionen, bei Hendrik Kayser für die vielen kleinen Hilfen und die gute Zeit in W2-2 256, bei Dr.-Ing. Thomas Rohdenburg ebenfalls für viele kleine Hilfen besonders bei der Richtungsschätzung, sowie bei Susanne Garre, Ingrid Wusowski, Anita Gorges und Frank Grunau für die Hilfsbereitschaft mit Formularen, Hörkabinen und Rechnern. Außerdem gilt mein Dank allen anderen Mitgliedern und Assoziierten der Arbeitsgruppe medizinische Physik und des internationalen Graduiertenkollegs Neurosensorik, die in vielerlei Hinsicht eine Unterstützung waren und mit denen ich auch außerhalb von Büro und Labor viel Spaß hatte.

Weiterer Dank geht an alle „Versuchspersonen“, die sich für zum Teil stundenlange psychoakustische Messungen in die kleine Hörkabine gesetzt haben.

Auch außerhalb der Arbeitsgruppe haben mich viele Menschen wissenschaftlich unterstützt und mir als Neuling einen unglaublich netten Einstieg in die binaurale Forschergemeinschaft bereitet. Dafür möchte ich mich bei Armin Kohlrausch, Steven van de Par, Tobias May, Bernhard Seeber, Georg Klump, Peter Heil, Berhard Laback, Tino Trahiotis, Les Bernstein, Richard Stern, Lutz Wiegrebe, Christoph Faller, Ifat Yasin, Torsten Marquardt und Emmanuelle Vincent bedanken.

Meiner Frau Ilka danke ich von ganzem Herzen für ihre grenzenlose Unterstützung, den unsagbar wertvollen Rückhalt sowie ihre vielen guten Ideen und Hilfen. Meinen Eltern danke ich ganz herzlich, dass sie immer für mich da waren, mein wissenschaftliches und technisches Interesse von meiner Kindheit an bestärkt haben und mein Studium so großzügig gefördert haben. Meinen Schwiegereltern danke ich ebenfalls für ihre Unterstützung und ihre erfahrenen Tipps. Meinem Freund Tim Lautenschläger danke ich für die tolle Zeit und die Unterkunft in Boston sowie für den wertvollen Austausch an Ideen, Gedanken und Erfahrungen. Auch meinen anderen lieben Freunden danke ich in gleicher Weise für ihre Unterstützung und den fortwährenden Erfahrungsaustausch.

Finanziell wurde diese Arbeit von der deutschen Forschungsgesellschaft durch das internationale Graduiertenkolleg „Neurosensorik“ und durch den Sonderforschungsbereich TRR 31 „Das aktive Gehör“ unterstützt.

Lebenslauf

Mathias Dietz

geboren am 19.06.1979 in Bad Wildungen

verheiratet seit dem 30.09.2006

Staatsangehörigkeit: deutsch

06/98	Gustav–Stresemann–Gymnasium in Bad Wildungen: Abitur
09/98 - 06/99	Fernmelderegiment 320 in Frankenberg Eder: Wehrpflicht.
04/99 - 09/99	TU Kaiserslautern: Teilzeit Fernstudienkurs Elektrotechnik
09/99 - 06/01	Volkswagen Coaching und Oskar–von–Miller–Schule in Kassel: Berufsausbildung zum Industrieelektroniker.
10/99 - 03/00	Fernuniversität Hagen: Teilzeit Fernstudienkurs Lineare Algebra I
10/01 - 09/03	Universität Münster: Grundstudium Physik.
09/03 - 03/04	University College London, UK: zwei Trimester Auslandsstudium.
04/04 - 07/06	Universität Münster: Hauptstudium Physik.
08/06 - 10/09	Universität Oldenburg, AG Medizinische Physik: Promotionsstudent im Sonderforschungsbereich „Das aktive Gehör“, sowie im internationalen Graduiertenkolleg „Neurosensorik“.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Dissertation hat weder in Teilen noch in ihrer Gesamtheit einer anderen wissenschaftlichen Hochschule zur Begutachtung in einem Promotionsverfahren vorgelegen. Teile der Dissertation wurden bereits veröffentlicht bzw. sind zur Veröffentlichung eingereicht, wie an den entsprechenden Stellen angegeben.

Oldenburg, den 8. September 2009