Honda Research Institute EU

Honda Research Institute Europe GmbH
https://www.honda-ri.de/

A Probabilistic Model for Binaural Sound Localization

Volker Willert, Julian Eggert, Jürgen Adamy, Raphael Stahl, Edgar Körner

2006

Preprint:

This is an accepted article published in IEEE Transactions on Systems, Man and Cybernetics - Part B. The final authenticated version is available online at: https://doi.org/[DOI not available]

A Probabilistic Model for Binaural Sound Localization

Volker Willert, Julian Eggert, Jürgen Adamy, Raphael Stahl, and Edgar Körner

***Abstract*—This paper proposes a biologically inspired and technically implemented sound localization system to robustly estimate the position of a sound source in the frontal azimuthal half-plane. For localization, binaural cues are extracted using cochleagrams generated by a cochlear model that serve as input to the system. The basic idea of the model is to separately measure interaural time differences and interaural level differences for a number of frequencies and process these measurements as a whole. This leads to two-dimensional frequency versus time-delay representations of binaural cues, so-called activity maps. A probabilistic evaluation is presented to estimate the position of a sound source over time based on these activity maps. Learned reference maps for different azimuthal positions are integrated into the computation to gain time-dependent discrete conditional probabilities. At every timestep these probabilities are combined over frequencies and binaural cues to estimate the sound source position. In addition, they are propagated over time to improve position estimation. This leads to a system that is able to localize audible signals, for example human speech signals, even in reverberating environments.**

***Index Terms*—Binaural hearing, probabilistic estimation, sound source localization.**

I. Introduction

THE PERCEPTION of our environment and interpersonal communication strongly depends on hearing. One of the primary abilities of the human auditory system is to localize sound sources in the environment. Sound localization serves as an important cognition feature, e.g., for attention control and self-orientation. Therefore, the development of a computational model of the human auditory localization process that is able to robustly localize real sound sources in natural environments is useful to improve the performance in many fields of application, for example, the interaction between humans and robots or binaural hearing aids for persons with hearing deficits.

There are three main requirements concerning sound localization: It would be desirable to: 1) accurately localize any kind of speech or sound source; 2) separate sound sources according to different positions; and 3) track moving sound sources. To deal with these problems the central auditory system of all mammals uses a common computational strategy [1], [2]. First, a variety of spatial localization cues is measured. Those cues that are from a single sound source are then grouped together and associated with the appropriate position in space. There are two kinds of spatial localization cues: monaural and binaural cues. Monaural cues are based on filter characteristics of the pinna of a single ear and are not addressed in this article. But there are two binaural (interaural) cues that are based on differences in timing and level of the sound at each of the two ears called interaural time differences (ITDs) and interaural level differences (ILDs).

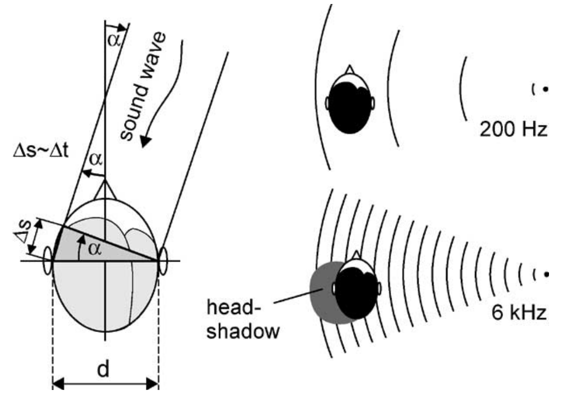

Fig. 1. Principles of the appearance of interaural time and interaural level differences (ITD and ILD). On the left the principle of time-delay between the two ears dependent on the sound source position angle is shown. On the right the principle of frequency- and position-dependent damping of the sound signal is illustrated, the so-called head-shadow effect.

Relating to Fig. 1, ITDs arise in the human auditory system because the two ears are spatially positioned with a distance d to each other given by the head dimensions. This leads to a difference in transmission time Δt of a signal arriving at the two ears. As a first approximation, this can be thought of as a difference in distance $\Delta s \propto \Delta t$ in the straight line from the sound source to the ears [3]. But this approximation is slightly inaccurate due to arising reflection and diffraction effects by the head, the shoulders, and the external ears, which lead to phase differences dependent on the frequency f [4]. Therefore, the measurement of ITDs is influenced by phase differences as well. Especially for wavelengths smaller than the head diameter ($f > 1500$ Hz) the distance Δs may be greater than one wavelength. This leads to an ambiguous situation where Δs does not correspond to a unique sound source position angle α [5]. For humans, there is no phase locking above 1500 Hz, and they are not sensitive to interaural phase shifts above that frequency [4], [6]. Furthermore, the time-delay Δt is not linearly mapped on α. The more the sound source moves to the side the slower Δt increases. This nonlinear dependence is approximately given by $\Delta t = d \sin(\alpha)/v$ following Fig. 1 with v being the acoustic velocity.

Manuscript received May 6, 2005; revised October 19, 2005 and January 4, 2006. This paper was reommended by Associate Editor P. K. Willett.

V. Willert, J. Adamy, and R. Stahl are with the Control Theory and Robotics Lab, Institute of Automatic Control, Darmstadt University of Technology, 64283 Darmstadt, Germany (e-mail: volker.willert@rtr.tu-darmstadt.de; adamy@rtr.tu-darmstadt.de; kpcr@gmx.de).

J. Eggert and E. Körner are with the Honda Research Institute, 63075 Offenbach, Germany (e-mail: julian.eggert@honda-ri.de; edgar.koerner@honda-ri.de).

Digital Object Identifier 10.1109/TSMCB.2006.872263

1083-4419/$20.00 © 2006 IEEE

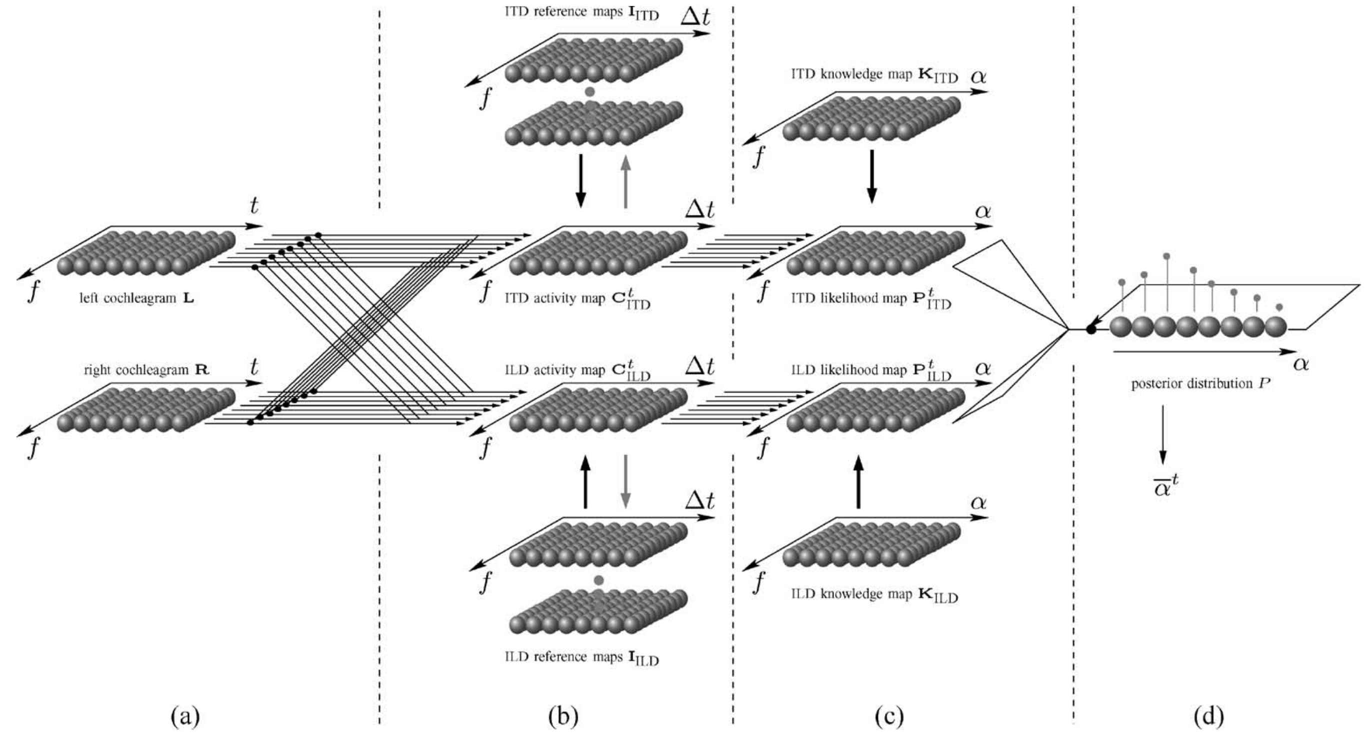
ITD reference maps $\mathbf{I}_{\text{ITD}}$
Δt
f
ITD knowledge map $\mathbf{K}_{\text{ITD}}$
α
t
Δt
α
f
left cochleagram $\mathbf{L}$
f
ITD activity map $\mathbf{C}^t_{\text{ITD}}$
f
ITD likelihood map $\mathbf{P}^t_{\text{ITD}}$
right cochleagram $\mathbf{R}$
t
ILD activity map $\mathbf{C}^t_{\text{ILD}}$
Δt
ILD likelihood map $\mathbf{P}^t_{\text{ILD}}$
α
α
posterior distribution P
f
f
f
$\overline{\alpha}^t$
Δt
α
f
f
ILD knowledge map $\mathbf{K}_{\text{ILD}}$
ILD reference maps $\mathbf{I}_{\text{ILD}}$
(a)
(b)
(c)
(d)

Fig. 2. System structure for binaural sound source localization used in this paper. The system mainly consists of four stages. (a) Measurement of discrete time t versus frequency f spectra for the left and right microphone of the robot head, explained in Section III. (b) Correlation-based extraction of binaural cues like ITD and ILD, represented with frequency f versus time-delay Δt matrices, so-called ITD/ILD activity maps, explained in Section IV. (c) Transformation of the activity maps to ITD/ILD likelihood maps that are frequency f versus position angle α matrices using learned ITD/ILD reference maps, explained in Section V. There is also the possibility to put in additional knowledge using ITD/ILD knowledge maps. (d) Combination of all the probability maps to one posterior distribution that is propagated over time t to estimate the most probable sound source position angle $\overline{\alpha}^t$ extracted from the posterior distribution, explained in Section VI.

By comparison, the ILD is strongly related to the loss of high frequencies caused by diffraction and refraction effects due to the geometry of the head and shoulders. Damping because of energy dissipation dependent on the material between the signal receptors also influences the ILD. In general, changing the azimuthal sound source position α towards the ipsilateral side (the side closest to the sound source) increases ILD by a decrease in the sound pressure on the contralateral hemifield (the side farthest to the sound source). This is not true anymore for source positions near $\pm 90^\circ$ due to the bright spot [7]. The wave recombines at the point opposite the incident angle, generating a local increase in intensity. Additionally, the level ratio of the acoustic pressures between ipsi- and contralateral ear increases with raising frequency f. The dependence between the level ratio and the frequency is highly nonlinear and different for every human or robot head. It can generally be observed that the higher the frequency the larger the level ratio between ipsi- and contra lateral ear [8]. This observation is called the head-shadow effect and is visualized in Fig. 1. The classic Duplex Theory proposes that high-frequency tones are localized by ILDs and low-frequency tones by ITDs so that the interaural cues are used exclusively each in the frequency range that produces unambiguous ITDs or ILDs [9].

The main focus of this paper is on modeling a biologically inspired system shown in Fig. 2 to estimate azimuthal sound source position in which the location estimate will be updated at every time step using previous information as well as the current measurement. The estimation of the sound source position is performed in a probabilistic way based on calculated cochleotopically[1] organized maps, so-called activity maps $\mathbf{C}$, which represent binaural cues, and learned representatives, so-called reference maps $\mathbf{I}$, which include learned binaural information for specific sound source positions. The system of Fig. 2 works as follows.

1) The input to the system are the so-called cochleagrams $\mathbf{R}$, $\mathbf{L}$ from the sound waves arriving at the left and right microphone achieved by a cochlea-simulating filterbank with a chosen number of frequency channels f_{c}. A cochleagram is a frequency versus time, i.e., f versus t, matrix and consists of amplitude values for every discrete timestep $t_{1:n}$ for every characteristic frequency $f_{\text{c}} = f_{1:m}$, with n being the maximal number of recorded timesteps and m being the number of frequency channels.
2) The cochleagrams are correlated in two different manners to get the activity maps $\mathbf{C}^t_{\text{ITD}}$ and $\mathbf{C}^t_{\text{ILD}}$ that comprise binaural information. The correlation-based extraction of binaural cues using ITD and ILD, is represented with two frequency versus time-delay, i.e., f versus Δt, matrices of the same format, which is particularly useful to handle ambiguities within the single ITD and ILD measurements and to provide a link between ILD and ITD. The activity maps $\mathbf{C}^t_{\text{ITD}}$ and $\mathbf{C}^t_{\text{ILD}}$ are compared with corresponding reference maps $\mathbf{I}_{\text{ITD}}$ and $\mathbf{I}_{\text{ILD}}$ that

[1] Cochleotopy: Spatial arrangement of the frequency-dependent sensitivity of the basilar membrane along the cochlea.

are learned for a number of position angles α and binaural cues c (ITD or ILD) to produce cue-dependent cochleotopically organized azimuthal maps, which we call likelihood maps. The integration of learned reference maps into the model enables the system to adapt to any kind of robot head without requiring an explicit model of the relation between the interaural cues and the azimuthal position angle.

3) The comparison of the maps with their references leads to two separate likelihood maps $\mathbf{P}^t_{\mathrm{ITD}}$ and $\mathbf{P}^t_{\mathrm{ILD}}$, one for the ITD and one for the ILD cue that contain discrete probability values for all correlation measures z^t for every cue c, every observed position angle α and every frequency channel f_{c} of the filterbank. The values of these likelihood maps are seen as a whole and interpreted as one single conditional probability, the likelihood $P(z^t|\alpha, f, c)$. This likelihood comprises the conditional probabilities $P(z^t|\alpha, f, c)$ that the correlation measurement z^t has happened at timestep t given the position angle α of the sound source, the frequency f, and the cue c. At every timestep the likelihood is combined with knowledge about the influence of the frequency f dependent on the cue c and position angle α and the influence of the cue dependent on the position angle. These influences are realized with frequency and cue dependent weightings memorized in the so called knowledge maps $\mathbf{K}_{\mathrm{ITD}}$ and $\mathbf{K}_{\mathrm{ILD}}$.
4) Using standard marginalization applied to $P(z^t|\alpha, f, c)$, a single-likelihood distribution $P(z^t|\alpha)$ only conditioned by the angle α is calculated as a result. The likelihood is propagated over time using a Bayesian approach to get the posterior $P(\alpha|z^t)$ from which the position angle estimate $\overline{\alpha}^t$ is extracted for every timestep t using standard estimation methods, e.g., the maximum aposteriori estimator [10]. The system is then able to estimate the position angle of the sound source and improve the estimation result over time.

In Section II, a short overview of the biological auditory pathway of sound source localization, which is the archetype of the system proposed here, is given. The Sections III–VI describe the different stages a)–d) of the presented localization system. Finally, Section VII shows some performance results.

II. Biological Archetype

The auditory signal processing can be divided into two consecutive parts: the transformation of the sound wave to spike trains and the representation and extraction of sound and speech information in the different auditory brain areas. Referring to Fig. 3, incoming sound waves at the pinna are first filtered in the outer ear that has filter characteristics with a bandpasslike transfer function. Then, the filtered sound waves activate oscillations on the eardrum that are transmitted via the middle ear to the cochlea in the inner ear. The transformation to spike trains representing neuronal activity is done by the haircells that are distributed on the basilar membrane, the primary organ of hearing. The basilar membrane acts like a filterbank by converting acoustic vibrations to a frequency dependent neuronal activity pattern on the cochlea. High-frequency oscillations activate haircells only at the entrance of the cochlea and the lower the frequency of the oscillations is, the more they expand into the cochlea onto the basilar membrane [3]. The first stage a) of our model in Fig. 2 models several mentioned properties of the basilar membrane explained in Section III. These mechanisms are very well researched and understood. However, the neuronal processing along the auditory pathway up to the cortex is still a heavily discussed subject in biological research.

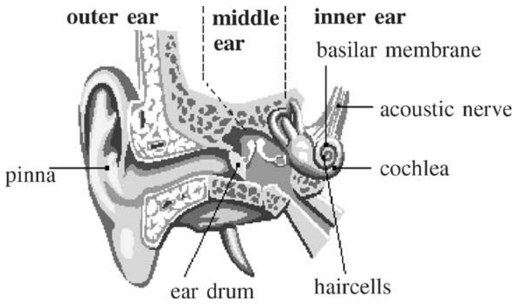

Fig. 3. Ear with the cochlea doing the transformation of the sound wave to neuronal activity.

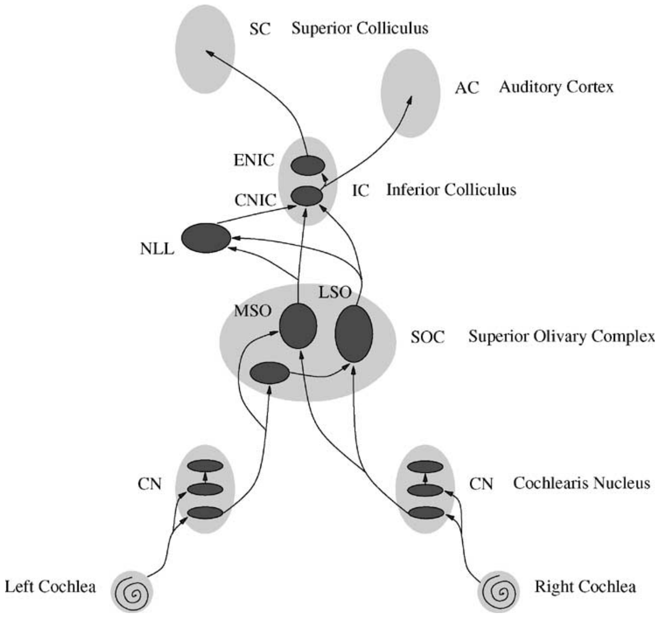

Fig. 4. Biological archetype for binaural localization. The different brain areas that contribute to the localization process with the required connections in the auditory pathway are shown.

The binaural pathway shown in Fig. 4 works to process location information about sound sources [11]. The cochleotopic mapping of the frequencies along the basilar membrane is passed through by the acoustic nerve to the next brain area, the cochlear nucleus (CN) [12]. The superior olivary complex (SOC) is the first place in the pathway where neurons receive input from both ears. The SOC comprises two main nuclei: The lateral superior olive (LSO) mainly works at high-frequencies using mainly sound amplitude differences and the medial superior olive (MSO) mainly works at low frequencies using time-delay differences [13]. Nevertheless, also some ITD-sensitive neurons are distributed across the LSO. Similar response types are found in neurons sensitive to ITDs in two signal types:

low-frequency sounds and envelopes of high-frequency sounds [13]. Studies in which the SOC is selectively damaged have demonstrated that it plays an essential role in the localization of the source of a sound [14]. The second stage b) of our model in Fig. 2 is explained in Section IV. It can be viewed as a technical model for the ILD and ITD extraction of LSO and MSO, that is, a model that provides the same functionality without attempting to be biologically faithful at the neuronal level.

One level higher, a tonotopic mapping according to different sound source positions can be found. In the inferior colliculus (IC), the same basic types of signal specificity remain, presumably due to inputs arising from the MSO and LSO [13]. However, other types of information gathered from the ear, such as timbre and frequency, appear to cross over at different points [14]. Neurons in the central nucleus of the IC (CNIC) are tuned to narrow frequency bands and also react to ITDs and ILDs [2]. The neurons of the external nucleus of the IC (ENIC) react differently compared to the neurons in the CNIC. Here, the neurons are tuned to wide frequency bands and single sound source positions [2]. It has been shown that lesions in the IC also result in a loss of the ability to localize sound [14]. The third stage c) of our model in Fig. 2 can be interpreted as a technical model that implements the frequency versus sound source position mapping of the ENIC and is explained in Section V.

The final decision of sound source position is believed to occur in auditory cortex (AC) and the combination of binaural cues with cues from other sensory inputs for combining sensory information is done in the superior colliculus (SC) [15]. The fourth stage d) of our model in Fig. 2 introduces a probabilistic model for SC to decide from where the sound has been sent and is explained in Section VI.

Some biologically inspired models of the MSO and LSO have already been developed, like the Jeffress Model [16], which extracts ITDs using a neuronal network with one layer, or the Stereausis Algorithm [17], [18], which extracts a two-dimensional (2-D) representation of ITDs using the phase shift of the filters of a cochlear filterbank. Several models for sound localization are based upon these principles and their extensions [8], [19]–[22].

Nevertheless, our model should be seen as a biologically motivated system with respect to the function and information mapping of the brain areas like MSO, LSO, as well as ENIC, and SC. However, it does not model the functionality with spiking neurons but with correlation-based methods and a functional probabilistic description for representation and estimation of lateral sound source position.

III. Cochlea Simulation

To model the cochlea and achieve a cochleotopic mapping of the signal frequencies over time, the auditory filterbank developed by Patterson *et al.* [23] is used. This filterbank is based on equivalent rectangular bandwidth (ERB)-filters, which are bandpass filters that have different bandwidths $b_{\mathbf{c}}$ for different characteristic frequencies $f_{\mathbf{c}} = \omega_{\mathrm{c}}/2\pi$. The Glasberg and Moore parameters [24] are chosen to gain a filterbank behavior comparable to the basilar membrane related to neuroacoustic experiments. Using these parameters one can compute f_{c} and b_{c} that lead to a typical overlap of the bandwidths on a logarithmic scale, as illustrated in Fig. 5. One ERB-filter approximates the frequency-dependent haircell activity at a particular location on the basilar membrane. For low frequencies the bandwidths are narrow with little overlap of neighboring filters while for higher frequencies the bandwidths get larger and the overlap increases. The ERB-filter $F_{\mathrm{c}}(z)$ formulated in (1) is a discrete eighth-order filter in z consisting of four Gammatone filters of second order as described in [25] with the corresponding z transform

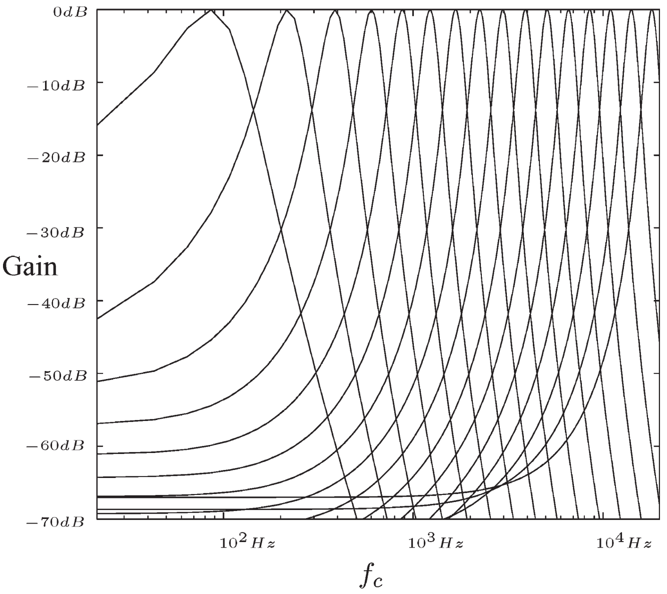

Fig. 5. Example of the gain responses of the Patterson–Holdsworth auditory filterbank with 16 characteristic frequencies. On a logarithmic scale the distance of neighboring characteristic frequencies f_{c} is approximately the same. Therefore, there is an increase of the overlap of the bandwidths with increasing frequency.

$$F_{\mathrm{c}}(z) = \left(\frac{Tz}{z^2 - 2e^{-b_{\mathrm{c}}T}\cos(\omega_{\mathrm{c}}T)z + e^{-2b_{\mathrm{c}}T}}\right)^4 \cdot \prod_{j,k=1}^{2}\left(z - e^{-b_{\mathrm{c}}T}\left(\cos(\omega_{\mathrm{c}}T) + (-1)^k \times \sqrt{3 + (-1)^j 2^{1.5}}\sin(\omega_{\mathrm{c}}T)\right)\right). \quad (1)$$

Every ERB-filter of the filterbank shifts the phase of its characteristic frequency $\omega_{\mathrm{c}} = 2\pi f_{\mathrm{c}}$ as can be seen in Fig. 6. Therefore, every characteristic frequency that is part of the input signal appears with a different phase distortion in the cochleagram. This complicates the process of comparing events across frequency channels because the frequency information of all frequency channels f_{c} at one timestep t in the cochleagram does not correspond to one single timestep of the input signal. To eliminate this disadvantage, time delays can be introduced to compensate the phase distortions. For convenience, we chose to use forward–backward filtering as described in [26] to achieve a phase-compensated cochleagram shown in Fig. 7 that is generated by stage a) of Fig. 2 from our system.

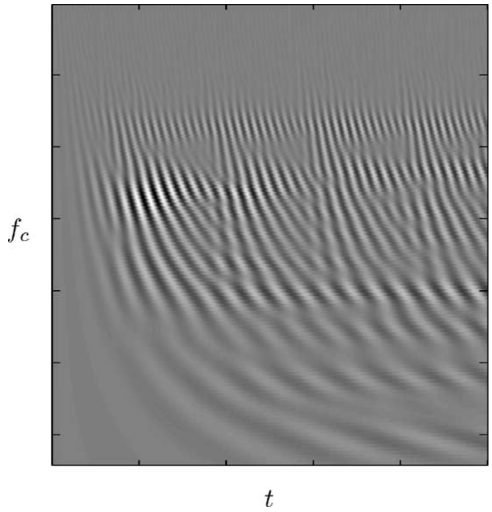

Fig. 6. Example of a cochleagram of a clap sound with the different phase shifts of the ERB filters.

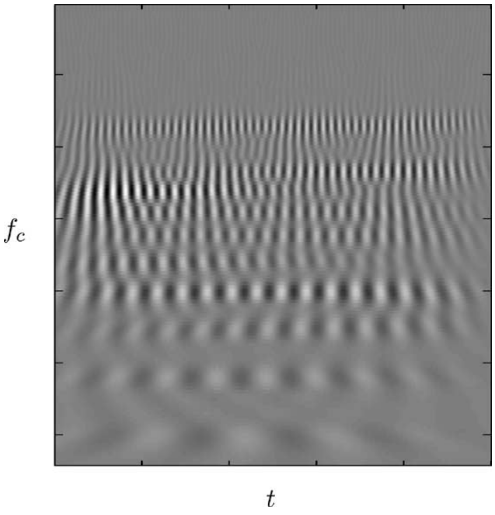

Fig. 7. Example of the phase-compensated cochleagram of Fig. 6.

Forward–backward filtering, that is, filtering the signal offline first forward and then backward in time results in precisely zero-phase distortion and doubles the filter order [26], [27]. This improves the ITD and ILD representation explained in Section IV because the different channels of the cochleagram show the activities of the frequencies as they arrive at the microphone without phase distortion. Therefore, the information of the frequency channels can be better combined at every timestep using small patches (time–frequency windows) of the cochleagrams of the left and right ear of a robot head. Nevertheless, forward–backward filtering is useful but not necessary for our system. It works also with the nonphase-compensated cochleagrams but the performance increases noticeably for the compensated system.

IV. Binaural Cue Extraction

The quality of the estimation of the azimuthal sound source position depends on the accuracy of the extraction of ITDs and ILDs. The difficulty in this context is that binaural differences in phase and level strongly vary over frequency (while the ITDs stay constant) and cannot be unambiguously assigned to the sound source position. For every frequency and position of a sound source there exists a characteristic ITD and ILD, but the measurement of the ITD and ILD cues is ambiguous. In our system, the cue extraction measures ITD and ILD activities for every possible time-delay and every frequency based on a straight correlation method. This leads to representations of ITDs and ILDs by 2-D maps over frequency versus time-delay (matrices), the activity maps $\mathbf{C}_{\mathrm{ITD}}$ and $\mathbf{C}_{\mathrm{ILD}}$ of Fig. 2, which can later be processed together to resolve the inherent ambiguities. Examples of such activity maps are shown in Figs. 8–11. The calculation of the activity maps is explained below.

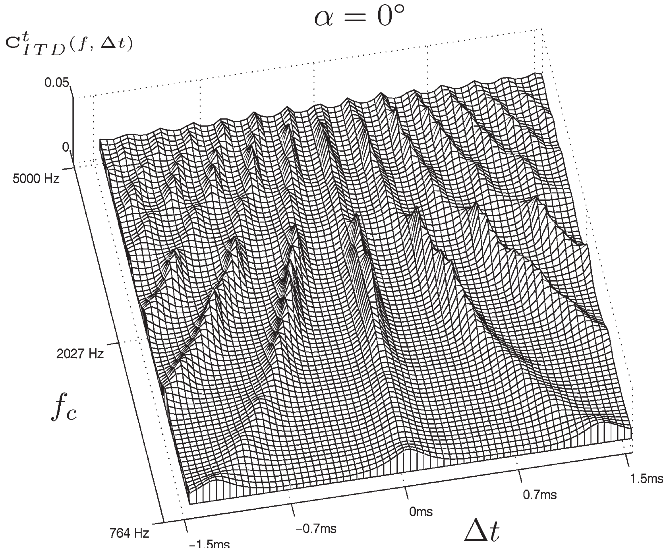

Fig. 8. Example of an ITD activity map, $\mathbf{C}_{\mathrm{ITD}}$, for $\alpha = 0°$ deviation of a clap sound.

In the rest of the section, we use the following notations: simple font for scalars (a, A) and bold for vectors and matrices ($\mathbf{a}$, $\mathbf{A}$). $\mathbf{1}$, $\mathbf{1}$ are a vector of ones and a matrix of ones, $\mathbf{A} \odot \mathbf{B}$ denotes a componentwise multiplication of two vectors or matrices and $\mathbf{A}^{\textcircled{\alpha}}$ a componentwise exponentiation by α of a vector or matrix.

To extract ITD and ILD activities for every timestep t small weighted patches of the left ear cochleagram $\mathbf{L} \in \mathbf{R}^{m \times n}$ are compared with weighted patches of the right ear cochleagram $\mathbf{R} \in \mathbf{R}^{m \times n}$. To do this, we define $\mathbf{W} \odot \mathbf{L}^{f,t}$ as a patch of signal amplitudes of the left cochleagram $\mathbf{L}$ anchored at the cochleagram position (f, t) of the size $k \times l$ with $k \ll m$ and $l \ll n$ and weighted with the window $\mathbf{W} \in \mathbf{R}^{k \times l}$ (e.g., a 2-D Gaussian window) that restricts the size of the patch and weights the position (f, t) the patch represents.

One possibility to measure activities for the ITDs is to assume that all amplitude values inside of a patch around (f, t) have a common time-delay between left and right ear, resulting in a time displacement Δt of the patch of the right cochleagram with respect to the patch of the left cochleagram. This assumption basically amounts to a search for correspondences of weighted cochleagram patches (displaced with respect to each other) $\mathbf{W} \odot \mathbf{L}^{f,t}$ and $\mathbf{W} \odot \mathbf{R}^{f,t+\Delta t}$.

To formulate the calculation of the ITD activities more precisely, we recur to a generative model [28] that is originally used in order to determine velocity distributions for optical flow computation and that can also be used here in order to approximately describe the connection between the cochleagrams of the left and right ear: every cochleagram patch $\mathbf{W} \odot \mathbf{L}^{f,t}$ is considered to be independent of its neighboring cochleagram patches $\mathbf{W} \odot \mathbf{L}^{f+\Delta f,t}$ at a particular time t. Furthermore, every cochleagram patch of the right ear $\mathbf{W} \odot \mathbf{R}^{f,t+\Delta t}$ is causally linked with a cochleagram patch of the left ear $\mathbf{W} \odot \mathbf{L}^{f,t,}$ in the following way: we assume that a patch of the left cochleagram $\mathbf{W} \odot \mathbf{L}^{f,t}$ within an associated time-delay Δt may be found in the right cochleagram $\mathbf{R}$ at position $(f, t + \Delta t)$, so that for this particular patch

$$\mathbf{W} \odot \mathbf{R}^{f,t+\Delta t} = \mathbf{W} \odot \mathbf{L}^{f,t} \tag{2}$$

holds. In addition, we assume that during this process the amplitudes are jittered by noise η, and that a mean shift and level difference of the amplitude of signal frequencies between the left and right cochleagram may occur. These variations in amplitude and bias are accounted for by an adjustable scaling parameter λ and an adjustable bias κ so that we arrive at

$$\mathbf{W} \odot \mathbf{R}^{f,t+\Delta t} = \lambda \mathbf{W} \odot \mathbf{L}^{f,t} + \kappa \mathbf{W} + \eta \mathbf{1}. \tag{3}$$

Solving (3) for $\eta\mathbf{1}$ and assuming that the image noise η is zero-mean Gaussian with variance σ_η, i.e.,

$$\eta\mathbf{1} = \mathbf{W} \odot \mathbf{R}^{f,t+\Delta t} - \lambda \mathbf{W} \odot \mathbf{L}^{f,t} - \kappa \mathbf{W} = \mathbf{A}$$
$$\rightarrow prob(\mathbf{A}) \sim e^{-\frac{1}{2\sigma_\eta^2}\|\mathbf{A}\|^2}$$

the determination that $\mathbf{L}^{f,t}$ is a match for $\mathbf{R}^{f,t+\Delta t}$ leads to the activity map elements

$$\mathbf{C}^t_{\text{ITD}}(f, \Delta t) = \frac{1}{\sigma_\eta\sqrt{2\pi}} e^{-\frac{1}{2\sigma_\eta^2}\left\|\mathbf{W}\odot(\mathbf{R}^{f,t+\Delta t} - \lambda\mathbf{L}^{f,t} - \kappa\mathbf{1})\right\|^2} \tag{4}$$

for the ITD cue containing the activities for all characteristic frequencies f and all time-delays Δt at a certain timestep t.

To gain a patch matching measure that is almost amplitude and bias invariant, we now proceed to make (4) independent of λ and κ. For this purpose, we maximize the measurement in (4) with respect to the scaling and shift parameters, λ and κ. This amounts to minimizing the exponent in (4), so that we want to find

$$\{\lambda^*, \kappa^*\} := \min_{\lambda,\kappa} \left\|\mathbf{W} \odot (\mathbf{R}^{f,t+\Delta t} - \lambda \mathbf{L}^{f,t} - \kappa \mathbf{1})\right\|^2. \tag{5}$$

Partially differentiating (5) with respect to λ and κ, setting these partial derivatives to zero and analytically solving the resulting equations

$$\frac{\partial}{\partial \lambda, \kappa} \left\|\mathbf{W} \odot (\mathbf{R}^{f,t+\Delta t} - \lambda \mathbf{L}^{f,t} - \kappa \mathbf{1})\right\|^2 = 0 \tag{6}$$

with respect to λ and κ leads to

$$\lambda^* = \frac{\varrho_{\mathbf{R}^{f,t+\Delta t},\mathbf{L}^{f,t}} \cdot \sigma_{\mathbf{R}^{f,t+\Delta t}}}{\sigma_{\mathbf{L}^{f,t}}} \quad \text{and} \tag{7a}$$

$$\kappa^* = \mu_{\mathbf{R}^{f,t+\Delta t}} - \lambda^* \cdot \mu_{\mathbf{L}^{f,t}} \quad \text{with} \tag{7b}$$

$$\mu_{\mathbf{X}} = \langle \mathbf{X} \rangle := \frac{\mathbf{1}^T \mathbf{X} \odot \mathbf{W}^{\textcircled{2}} \mathbf{1}}{\mathbf{1}^T \mathbf{W}^{\textcircled{2}} \mathbf{1}} \tag{7c}$$

$$\sigma^2_{\mathbf{X}} = \langle \mathbf{X}^{\textcircled{2}} \rangle - \langle \mathbf{X} \rangle^2 \quad \text{and} \tag{7d}$$

$$\varrho_{\mathbf{X},\mathbf{Y}} = \frac{1}{\sigma_{\mathbf{X}} \cdot \sigma_{\mathbf{Y}}} \langle (\mathbf{X} - \mu_{\mathbf{X}}\mathbf{1}) \odot (\mathbf{Y} - \mu_{\mathbf{Y}}\mathbf{1}) \rangle \tag{7e}$$

where $\mathbf{X}$, $\mathbf{Y}$ have to be replaced by $\mathbf{R}^{f,t+\Delta t}$, $\mathbf{L}^{f,t}$, respectively. The weighted empirical correlation coefficient $\varrho_{\mathbf{R}^{f,t+\Delta t},\mathbf{L}^{f,t}} \in [-1; 1]$ is an efficient standard correlation measure [29] with 1 meaning the patterns $\mathbf{R}^{f,t+\Delta t}$ and $\mathbf{L}^{f,t}$ are fully correlated and -1 the patterns are fully anticorrelated.

Inserting (7a) and (7b) into (4), so that $\lambda = \lambda^*$ and $\kappa = \kappa^*$, leads to the final ITD map activities

$$\tilde{\mathbf{C}}^t_{\text{ITD}}(f, \Delta t) = \frac{1}{\sigma_\eta\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\left(\frac{\sigma_{\mathbf{R}^{f,t+\Delta t}}}{\sigma_\eta}\right)^2\left(1-\sigma^2_{\mathbf{L}^{f,t},\mathbf{R}^{f,t+\Delta t}}\right)}. \tag{8a}$$

Additionally, the activities of the ITD map $\tilde{\mathbf{C}}^t_{\text{ITD}}(f, \Delta t)$ are normalized with respect to the time-delays Δt in the way

$$\mathbf{C}^t_{\text{ITD}}(f, \Delta t) = \frac{\tilde{C}^t_{\text{ITD}}(f, \Delta t)}{\sum_{\Delta t} \tilde{\mathbf{C}}^t_{\text{ITD}}(f, \Delta t)}. \tag{8b}$$

This enhances the activities of a unimodal distribution over Δt for a fixed characteristic frequency, and reduces the activities with multimodal or flat distributions within a frequency channel.

Equation (8a) ensures that local changes in level or bias have minimal influence on the accuracy of the correlation. The binaural signals contribute to the ITD activity map primarily when the signal-to-noise ratio is high. In all experiments the variance of the noise σ^2_η is chosen to be 10% of the mean variance of all local patches from the left and right cochleagrams.

The calculation of the ILD activity map $\mathbf{C}^t_{\text{ILD}}(f, \Delta t)$ arises from (7a) that describes the optimal λ^* for the compensation of level differences between the two ears. The ratio of the variances $\sigma_{\mathbf{L}^{f,t}}/\sigma_{\mathbf{R}^{f,t+\Delta t}}$ is a measure for the level differences and the multiplication with the correlation coefficient $\varrho_{\mathbf{R}^{f,t+\Delta t},\mathbf{L}^{f,t}}$ causes the level ratio to contribute to the correct time-delay.

For the ITD measurement according to (8), the level ratio (and therefore the ILD) is explicitly neglected using the correlation coefficient because it is normalized due to the means and the variances. For the ILD measurement, we proceed in a complementary way using only the level ratio and neglecting the correlation coefficient, that is, neglecting the ITD. To get the ratio values in a proper range, the logarithmic ratio is

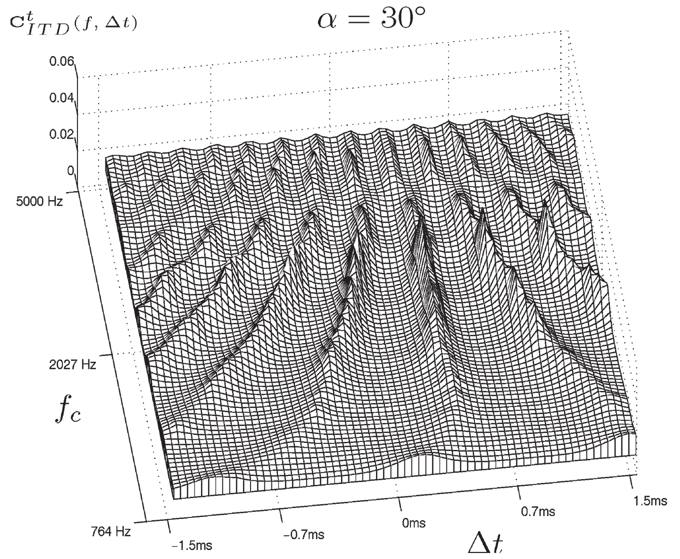

Fig. 9. Example of an ITD activity map, $\mathbf{C}_{\mathrm{ITD}}$, for $\alpha = 30°$ deviation of the same clap sound as in Fig. 8.

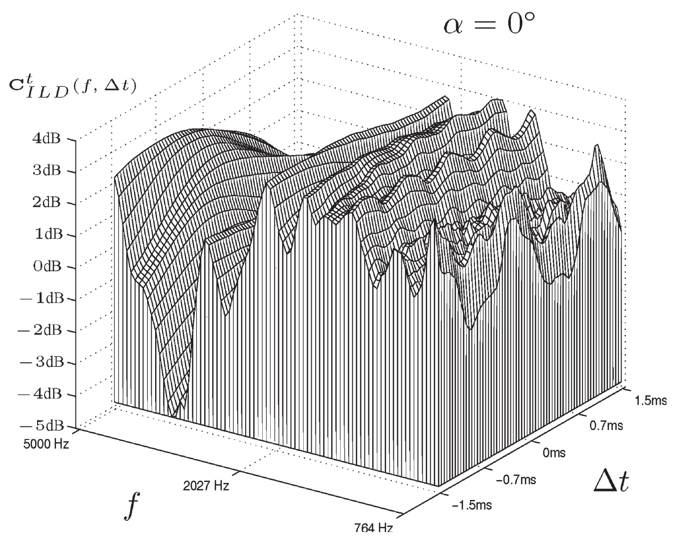

Fig. 10. Example of an ILD activity map, $\mathbf{C}_{\mathrm{ILD}}$, for $\alpha = 0°$ deviation of the same clap sound as in Fig. 8.

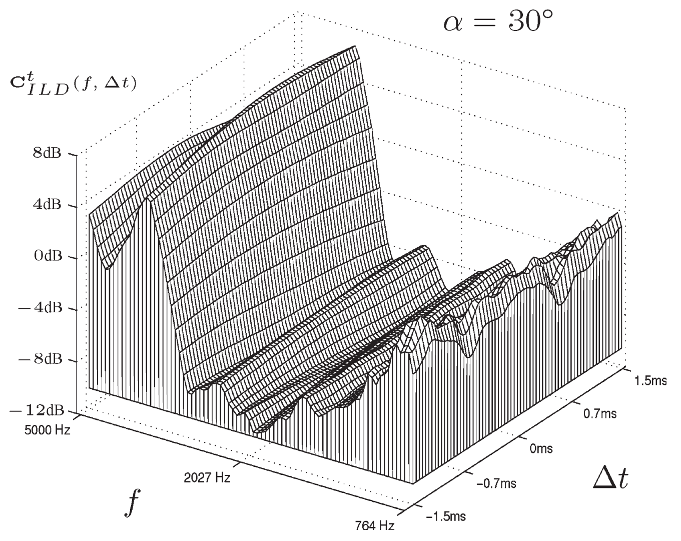

Fig. 11. Example of an ILD activity map $\mathbf{C}_{\mathrm{ILD}}$ for $\alpha = 30°$ deviation of the same clap sound as in Fig. 8.

calculated. This modified version of λ^* serves as the ILD activity formulation

$$\mathbf{C}^t_{\mathrm{ILD}}(f, \Delta t) = 20 \log_{10} \frac{\sigma_{\mathbf{L}^{f,t}}}{\sigma_{\mathbf{R}^{f,t+\Delta t}}}. \tag{9}$$

The level ratio is computed for all inspected time-delays Δt to be sure the correct level ratio corresponding to the real azimuthal position is existent in the ILD measurement. This leads to an ILD activity map with the same dimensions as the ITD activity map measured for all frequencies and time-delays. In Figs. 8 and 9, examples of ITD activity maps that correspond to two different azimuthal positions $\alpha = 0°$ and $\alpha = 30°$ for a binaural clap sound[2] are given. The two maps are differently activated with approximately the same periodic and symmetric pattern but differently shifted in time Δt dependent on the position angle α. As the sound source relocates to the side $\alpha = 30°$, the activities in Fig. 9 are shifted approximately about 0.5 ms compared to Fig. 8. In Figs. 10 and 11, examples of ILD activity maps that correspond to the examples of the ITD activity maps in Figs. 8 and 9 are given. Notice that for $\alpha = 0°$ the activities are in the range of $[-5\text{ dB}; 4\text{ dB}]$ and vary strongly over frequency. For $\alpha = 30°$ the activities are in the range of $[-10\text{ dB}; 8\text{ dB}]$ and the profile along the frequency axis in Fig. 10 is completely different as compared to Fig. 11. Interestingly, the activities of the ILD map along the time-delay axis (e.g., in Fig. 10) are often higher at time-delays Δt where there is high activity in the corresponding ITD map (e.g., in Fig. 8) as well. At each timestep these two maps jointly characterize the sound source position angle.

To achieve a probabilistic description of the sound source position, as shown in part c) of the system in Fig. 2, reference maps $\mathbf{I}_{\mathrm{ITD}}$ and $\mathbf{I}_{\mathrm{ILD}}$ for representative positions around the robot head are learned (see part b) of the system structure in Fig. 2). The use of reference maps that need to be learned anew for different robot heads enables the system to adapt to any kind of robot head without the necessity to rely on functions that explicitly describe the relation between the interaural cues and the azimuthal position. These head-dependent relations are therefore implicitly contained in the reference maps.

Learning is done in a supervised way, which means that the sound position is known during the learning process. The learning step is done separately for the ITD and for the ILD measurements and results in ITD and ILD reference maps $\mathbf{I}_{\mathrm{ITD}}(\alpha, f, \Delta t)$, $\mathbf{I}_{\mathrm{ILD}}(\alpha, f, \Delta t)$ that are representatives for specific positions α for all frequencies f of the cochleagram.

Fig. 12 illustrates how learning is done in practice. A sound source S is placed at a particular distance D and azimuthal position α. Then, different signals, which consist of a variety of wideband sound signals, e.g., sentences and sounds spoken or generated by different persons, and sinusoidal narrowband test

[2] All signals were emitted by a loudspeaker and recorded by the microphones of a robot head in an anechoic chamber (see also Fig. 12).

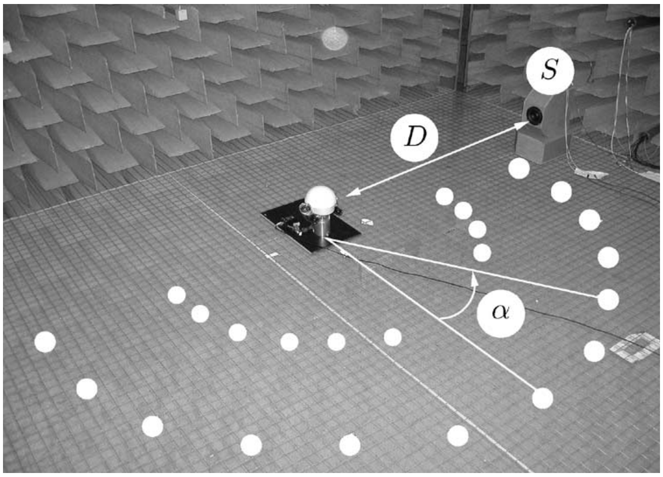

Fig. 12. Experimental setup for learning reference maps for a specific robot head (Miiro from Darmstadt University of Technology) in an anechoic chamber. The microphones (low-cost/no-name dynamical cardioid microphones) are placed on an aluminum bar with a distance of 14 cm and a spherical massive styrofoam head in between. The signals are recorded with a standard PC soundcard (Creative Soundblaster AWE64) with a sampling rate of 44 100 Hz and a 16-bit A/D converter.

signals, e.g., generated from a signal generator, covering the frequency spectrum that the system should be able to process, are emitted by a loudspeaker in an anechoic chamber and the corresponding ITD and ILD activity maps $\mathbf{C}_{\mathrm{ITD}}$ and $\mathbf{C}_{\mathrm{ILD}}$ are calculated. All activity maps for the same position are averaged over the number of measurement timesteps per position t_α according to

$$\mathbf{I}_{\mathrm{c}}(\alpha, f, \Delta t) = \frac{1}{t_\alpha} \sum_{t_\alpha} \mathbf{C}_{\mathrm{c}}^{t_\alpha}(f, \Delta t) \tag{10}$$

to gain head-specific and separate ITD and ILD reference maps $\mathbf{I}_{\mathrm{c}}$ with $c \in \{\mathrm{ITD}, \mathrm{ILD}\}$ that are general concerning the bandwidth and type of signal. This procedure is done for all chosen positions α. Because the 2-D reference maps change their patterns smoothly over azimuthal angle, only a few positions for learning have to be processed. For our application, thirteen positions for learning reference maps have been chosen that are equally distributed over azimuthal plane in a range of $\alpha \in [-90°; 90°]$ according to the gaze direction of the robot head with a common distance D of 2 m. The distance D is not a critical parameter as long as it is chosen to be larger than 1 m. For the classification results in Section VII, the sound sources have been placed at different positions in the range of 1–3 m.

V. Likelihood Formulation

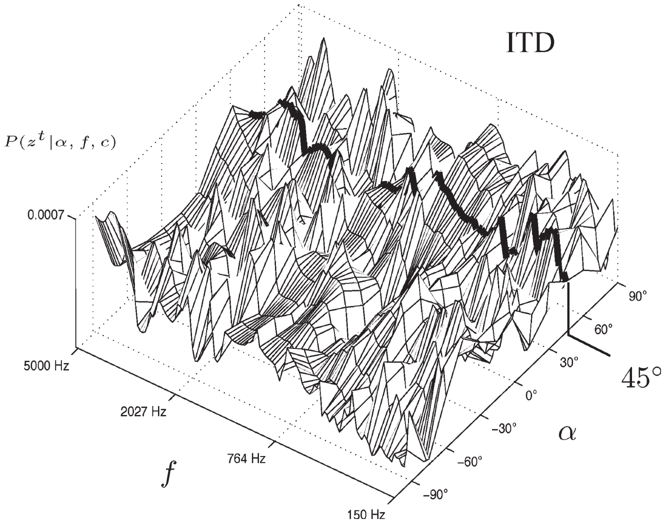

Fig. 13. Example of a likelihood distribution only measuring the ITD cue $P(z^t|\alpha, f, c = \mathrm{ITD})$ for 45°.

After learning, the system is able to calculate a likelihood distribution for sound source position-estimation in the following way: at every timestep t the binaural signal is preprocessed with the cochlear filterbank and ITD and ILD correlation measurements are calculated using (8a) and (9). The resulting ITD and ILD activity maps $\mathbf{C}_{\mathrm{c}}^t(f, \Delta t)$ are then compared with learned reference maps $\mathbf{I}_{\mathrm{c}}(\alpha, f, \Delta t)$ to gain one single discrete probability distribution. This single distribution is represented in the two likelihood maps $\mathbf{P}_{\mathrm{ITD/ILD}} := P(z^t|\alpha, f, c)$ that are calculated as follows:

$$P(z^t|\alpha, f, c) = \frac{1}{\sigma_P\sqrt{2\pi}} e^{\frac{1}{2\sigma_P^2} \sum_{\Delta t}(\mathbf{C}_{\mathrm{c}}^t(f, \Delta t) - \mathbf{I}_{\mathrm{c}}(\alpha, f, \Delta t))^2}. \tag{11}$$

Every distribution describes the conditional probability that the measurement z^t has happened at timestep t given a sound source position α for each frequency channel f and each binaural cue c. The measurements z^t are the comparisons between the activity maps $\mathbf{C}_{\mathrm{c}}^t(f, \Delta t)$ and the corresponding reference maps $\mathbf{I}_{\mathrm{c}}(\alpha, f, \Delta t)$ by the sum of squared differences summed over all time-delays Δt measured at every timestep t for every frequency f for both cues c and for all position angles α for which reference maps have been learned. In all experiments the standard deviation σ_P is chosen channelwise to be 50% of the mean value over α of all sums of squared differences between $\mathbf{C}_{\mathrm{c}}^t(f, \Delta t)$ and $\mathbf{I}_{\mathrm{c}}(\alpha, f, \Delta t)$.

In Figs. 13 and 14, the likelihood distribution separated into ITD and ILD probabilities is shown. The speech signal is sent from 45°. The correct position is marked with a bold black line within the distribution to show which frequencies contribute to the position and how strong the contribution is. It can be seen that the probability values along the frequency axis for position 45° are not outstanding. To enable a correct estimation of the sound source position, we want to get rid of the dependencies of the measurement z^t on the frequency f and the cue c to arrive at a conditional probability distribution $P(z^t|\alpha)$ that only depends on the sound position α. Furthermore, we want to estimate the position for every timestep t and propagate the estimate over time.

A. Estimation

As described in Section I, the localization ability differs between the cues c dependent on the frequency f and the localization angle α. This is caused by the geometry and material

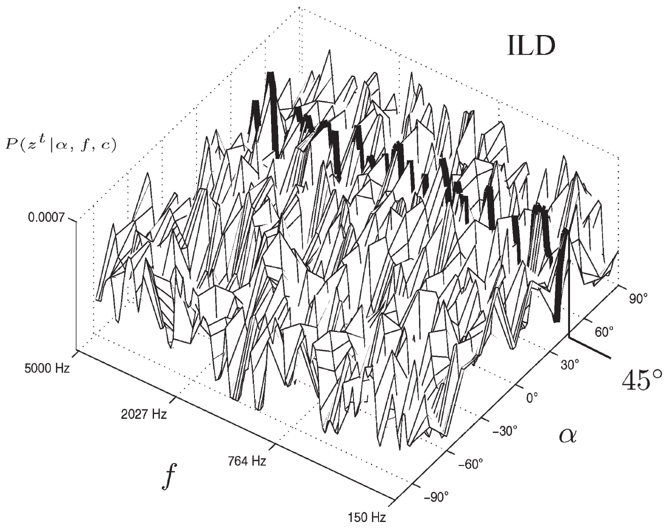

Fig. 14. Example of a likelihood distribution only measuring the ILD cue $P(z^t|\alpha, f, c = \text{ILD})$ for 45°.

of the head and leads to ambiguities in the likelihood distribution (multiple peaks). To reduce the ambiguities, conditional probability distributions $P(f|\alpha, c)$ and $P(c|\alpha)$ can be defined that incorporate knowledge about frequency and position angle dependent preference of ITD or ILD cue for sound localization.

Concerning the Duplex Theory [9], the probability values of $P(f|\alpha, c)$ for the ITD cue c can be given a smaller weight for high frequencies because the ambiguity of the ITD measurement rises with higher frequencies. Similarly, the probability values of $P(f|\alpha, c)$ for the ILD cue c can be biased for high frequencies because there is more binaural level difference expected.

Due to the worse resolution of time-delay Δt of the ITD cue c with increasing azimuthal angle α of the sound source, the probability values of $P(c|\alpha)$ for the ITD cue c can be chosen so as to contribute less for large azimuthal angles. In contrast to the ITD weighting, the probability values of $P(c|\alpha)$ for the ILD cue c can be given a smaller weight for small azimuthal angles because in this case the binaural level differences do not change significantly (for the measurements on the robot head we used).

The probability values of $P(f|\alpha, c)$ and $P(c|\alpha)$ are stored in the knowledge maps $\mathbf{K}_{\text{ITD}}$ and $\mathbf{K}_{\text{ILD}}$ of the localization system shown in part c) of Fig. 2. They serve as separate frequency and cue weightings conditioned by the cue and angle, respectively. If there is no need to consider such preferences the probabilities are assumed to be equally distributed.

Combining all the conditional probability distributions $P(z^t|\alpha, f, c)$, $P(f|\alpha, c)$, and $P(c|\alpha)$ by applying marginalization [30] leads to the final likelihood distribution

$$P(z^t|\alpha) = \sum_{f,c} P(z^t|\alpha, f, c) \cdot P(f|\alpha, c) \cdot P(c|\alpha). \tag{12}$$

VI. Position Estimation

To gain an estimate for the sound source position $\overline{\alpha}^t$ at every timestep t, as outlined in part d) of Fig. 2, the posterior distribution $P(\alpha|z^t)$ over α has to be calculated. This is done using Bayes' Theorem [30] that is applied to the marginalized likelihood $P(z^t|\alpha)$ from (12) putting in prior knowledge $P(\alpha)$ about the preferential treatment of sound positions α in the following way:

$$P(\alpha|z^t) \propto P(z^t|\alpha) \cdot P(\alpha). \tag{13}$$

An assumption has to be made according to the prior $P(\alpha)$. An appropriate prior can be chosen if general knowledge is available about the environment the robot head is working in or the task for which sound localization is used. A possible example is a discussion scenario with several people. The prior can be chosen to: 1) restrict the localization task to the visual field to focus on several speaking people the robot is looking at or 2) to emphasize to regions outside its visual field to react to speakers the robot is currently not looking at.

The final sound source position estimation $\overline{\alpha}^t$ is calculated using a maximum *a posteriori* (MAP) estimation [10]

$$\alpha_{\text{MAP}}^{-t} = \text{argmax}_{\alpha} \left(P(\alpha|z^t) \right). \tag{14}$$

A. Propagation

To circumvent the problem of making prior assumptions, the prior $P(\alpha)$ can be formulated using the previous posterior, which means involving previous estimates in the prior assumptions and propagating the estimates over time. The simplest assumption that can be made is that within the time interval of propagation no changes in the sound source position appear. This leads to the following time-dependent predictive prior:

$$P(\alpha) := P^t(\alpha) = P(\alpha|z^{t-\Delta t}) \tag{15}$$

whereas the momentary prior is simply the previous posterior. Such a predictive model can very easily be augmented to incorporate movements of the sound source. To consider slight changes we extend the predictive prior to be a smoothed version of the previous posterior leading to the following prior formulation:

$$P^t(\alpha) = \sum_{\alpha'} P(\alpha|\alpha') \cdot P(\alpha'|z^{t-\Delta t}) \tag{16}$$

with $P(\alpha|\alpha')$ chosen to be Gaussian distributed. This has the effect that the larger the variance of $P(\alpha|\alpha')$ the less influence the prior has on the new likelihood.

Fig. 15 shows four posteriors estimated at four different consecutive times using (13) and an equally distributed prior $P(\alpha)$ for a speech signal located at azimuthal angle $\alpha = 45°$. Since the sound source is not moving and the points in time are close to each other, processing only very short parts of the signal, the posteriors look very similar. Using (16), the prior $P^t(\alpha)$ becomes time-dependent and the posteriors are combined over time t. The result is shown in Fig. 16. The posterior distribution gets more and more sharpened and peaked at the correct position. That means that the probability that $\alpha = 45°$ is the correct position of the sound source gets larger with increasing time t.

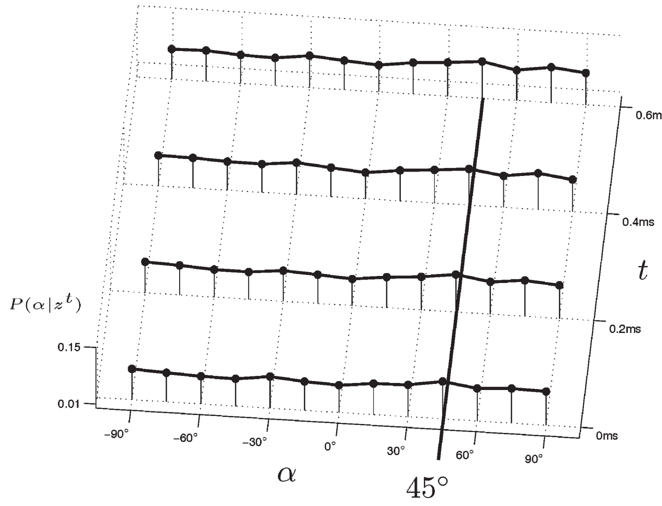

Fig. 15. Example of four posteriors $P(\alpha|z^t)$ estimated at different times for $\alpha = 45^\circ$. Since the sound source does not move and the time between the measurements is very short the posteriors are more or less the same. The highest probability value is at the correct position but not outstanding.

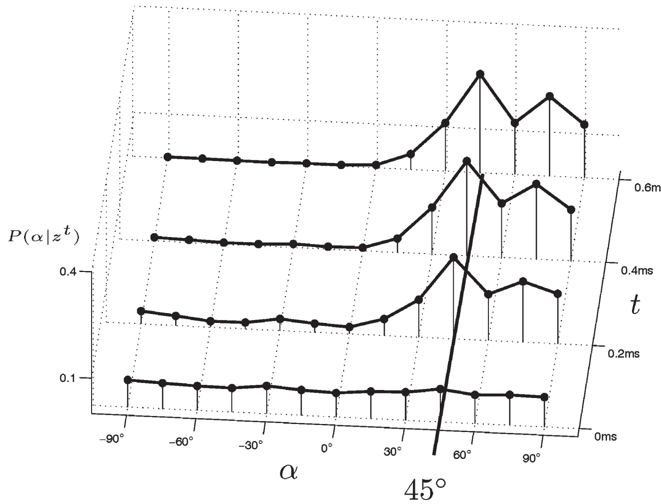

Fig. 16. Example of four posterior probabilities $P(\alpha|z^t)$ with the prior $P^t(\alpha)$ chosen to be propagated over time using (16) for the same signal as in Fig. 15.

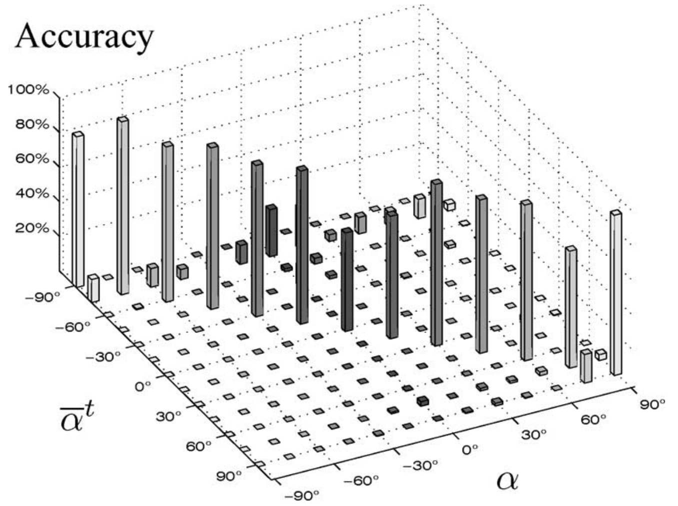

Fig. 17. Classification results for ITD cue only for signals recorded in an anechoic chamber resulting in an accuracy of 86.9%.

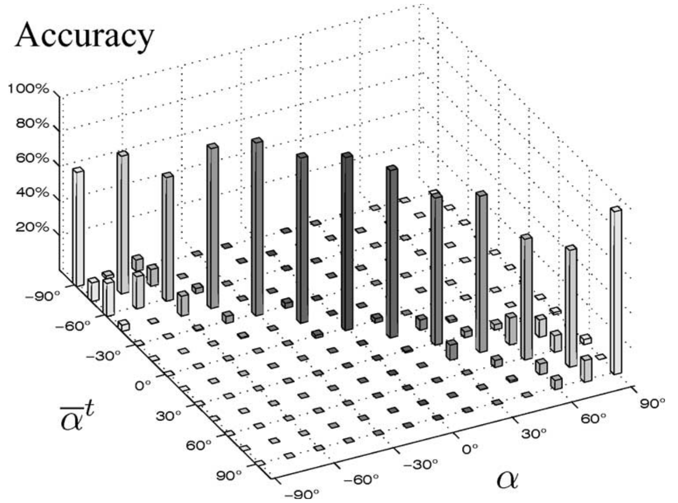

Fig. 18. Classification results for ILD cue only for signals recorded in an anechoic chamber resulting in an accuracy of 89.1%.

VII. Classification Results and Accuracy

The training of the reference maps and the classification of the test speech signals are carried out with a learn- and test-set of speech signals covering different types of male and female speakers saying different sentences. This is done for 13 angle positions equally distributed over the azimuthal range in the field of view of the robot ranging from $\alpha = -90^\circ$ up to $\alpha = 90^\circ$ according to the frontal sound occurrence, with 15° resolution between adjacent positions and discrete distances of 1, 2, and 3 m.

The training of the reference maps (see also Section IV) was done once in an anechoic chamber in which 13 reference maps each for ITD and ILD have been trained using 13×100 training signals for 13 different positions and 100 speech and sound signals.

The classification results for the test sets recorded in the anechoic chamber are based on different signal parts at $13 \times 38 \times 47$ different times whereas for each of the 13 classes 38 speech and sound signals are processed over 47 timesteps t for all three distances 1, 2, and 3 m. Every timestep lasts 0.023 ms.[3] The test sets recorded in the reverberant room were spoken from a male person in which the test sets of the anechoic chamber were emitted from a loudspeaker. There is no knowledge induced to consider preferences on some positions or cues, which means $P(f|\alpha, c)$ and $P(c|\alpha)$ are equally distributed.

A. Anechoic Chamber

Analyzing only the ITD cue leads to the classification results shown in Fig. 17 resulting in an accuracy of 86.9%. The confusion matrix is displayed, with $\overline{\alpha}^t$ being the estimated position angle and α the actual position angle. For the ILD cue only the confusion matrix is shown in Fig. 18 resulting in

[3] Because the sampling rate was chosen 44 100 Hz one timestep lasts 0.023 ms $\approx$ (44 100 Hz)$^{-1}$.

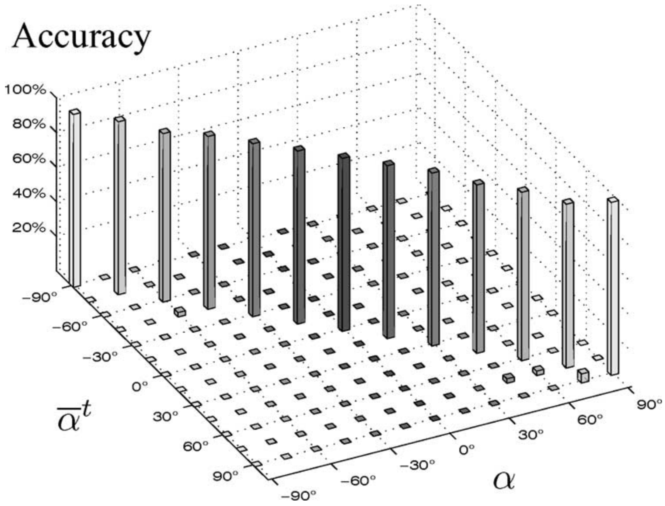

Fig. 19. Classification results for joint classification of ITD and ILD cues for signals recorded in an anechoic chamber resulting in an accuracy of 98.9%.

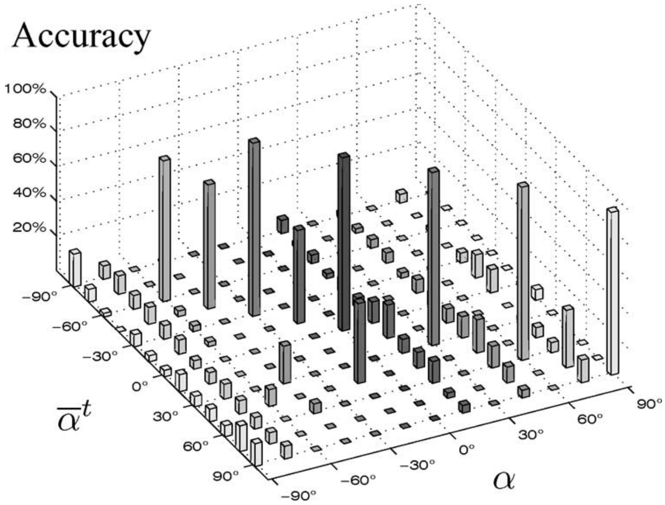

Fig. 20. Classification results for the classification of ITD cue only for signals recorded in a reverberant room and integrated over 0.2 ms after onset resulting in an accuracy of 61.7% for a location resolution of 15° and 84.9% for a location resolution of 30°.

an accuracy of 89.1%. The results gained by combining both cues and propagating the distributions over time can be seen in Fig. 19 resulting in an overall classification result with an accuracy of 98.9% for a location resolution of 15°.

B. Reverberant Room

In Figs. 20–22, the separated (ITD and ILD) and joint classification results for a male person saying *Hallo Miiro* (the name of the robot) are shown using the same head and the unmodified system with the reference maps trained in the anechoic chamber. The reverberant room was a normal living room with a reverberation time $T_{60} \approx 0.43$ s, a size of $5 \times 5 \times 2.50$ m and with the first early reflections arriving at ≈ 0.14 ms. Only the first 0.2 ms after onset of the now reverberated signal (which equals only the part *all* of *Hallo*) are processed to reduce the errors because of sound reflections. The onset is detected by applying an experimentally chosen threshold dependent on every frequency channel and starting the integration when the amplitude in one of the channels has reached its threshold (therefore the H of *Hallo* is neglected). The result is still reasonable with an accuracy of 61.7% for ITD, 33.2% for ILD, and 68.7% for the joint classification of both cues for a location resolution of 15° and an accuracy of 87.9% for both cues and a location resolution of 30°.

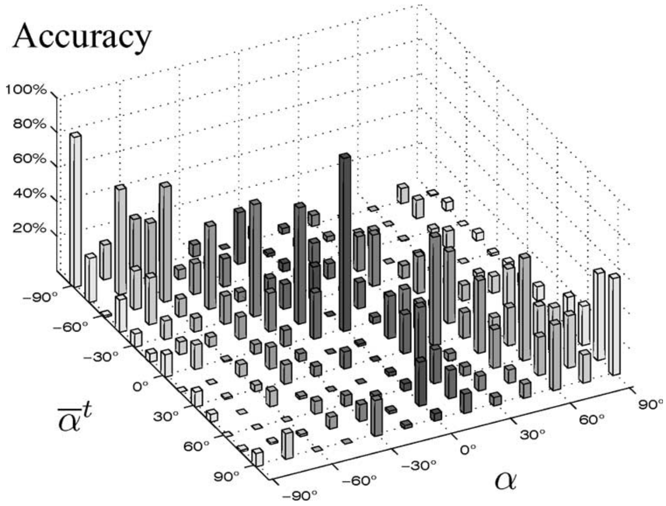

Fig. 21. Classification results for the classification of ILD cue only for signals recorded in a reverberant room and integrated over 0.2 ms after onset resulting in an accuracy of 33.2% for a location resolution of 15° and 39.6% for a location resolution of 30°.

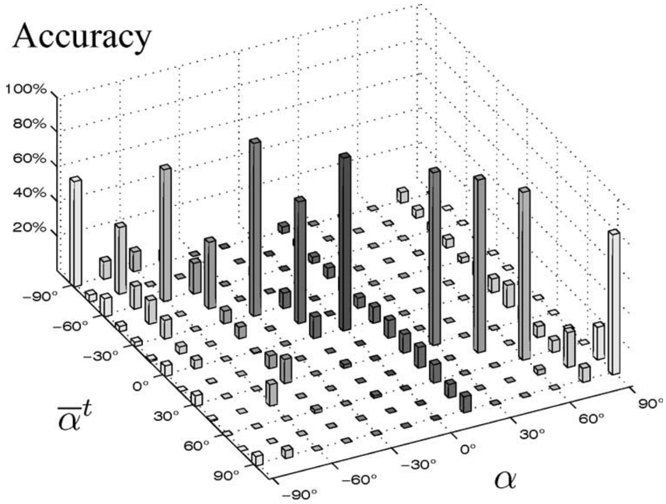

Fig. 22. Classification results for joint classification of ITD and ILD cues for signals recorded in a reverberant room and integrated over 0.2 ms after onset resulting in an accuracy of 68.69% for a location resolution of 15° and 87.9% for a location resolution of 30°.

VIII. Discussion and Outlook

We see that the proper separation of ITD and ILD cues extracted at the beginning of the localization process and the final combination of position hypotheses for classification at the end of the localization process improves the sound source positioning capabilities considerably. In general, we have observed that with rising resolution in the locations for the classification task the discrimination ability is reduced because confusion between neighboring positions occurs more often.

With the rising number of reflected sound waves occurring at the two microphones the ambiguities in the probability distribution rise and the probability values for every location go down and more or less equalize. This leads to a degradation of the performance in the reverberant environment. The number of timesteps used for integrating the information influences the performance only in a reverberant environment. With the experimentally chosen 0.2 ms we get enough information to integrate and resolve ambiguities in the measurements but also do not integrate too long so that the reflections do not introduce enough disturbed measurements that degrade the classification results. Even better performance can be obtained by using more than just one onset of a signal but this has not yet been studied within our research.

The ability of humans to localize sound sources in reverberant spaces very well is thought to be assisted by the precedence effect, which assumes that the perception of interaural cues provided by an indirect (reflected) sound is strongly influenced by the presence of the direct sound [31]. This implies the need for a method that is able to integrate as much interaural information as possible before the first reflection reaches the ears. There are some researchers trying to analyze and model these effects [31]–[33].

Beside the estimation and compensation of incoming echo [33] there is an interesting paper by Faller and Merimaa [32] who try to model some aspects of the precedence effect and are able to alternately localize concurrently active sources while ignoring the reflections and superposition effects. The main idea is to select ITD and ILD cues only from critical bands with high energy using an interaural coherence measurement (ICM) for selection. The ICM is the maximum value of the normalized cross correlation between the interaural signals.

Compared to our system, the probability value of the MAP estimator is similar to the ICM. Therefore, it would be easy to introduce the same selection mechanism they use to our system by replacing the ICM with the probability value of the MAP estimator. However, they do not consider the way the auditory system combines information from different critical bands. Further on, the method for selecting or discarding binaural cues is threshold-based and the localization process is purely done on the maximum values of the cross correlation, which means only one hypothesis for each cue per timestep is processed.

Our system is able to handle ambiguous and weak measurements by keeping several position hypotheses concurrently and improve or resolve ambiguous estimations using a proper integration method. Instead of a pure selection mechanism with our system, a similar processing with a slightly smoother transition could be implemented with adaptive weights in the knowledge maps. This but also localization and tracking of several speakers will be part of our future research.

IX. Conclusion

We have presented a binaural localization system that is able to localize human speech signals in the azimuthal plane with a very high accuracy of up to 98.9%. The modeling of the system is inspired by the human auditory pathway for sound localization and keeps several aspects of biologically equivalent tonotopic mapping of ITD and ILD as well as frequency versus sound-location mapping. From the point of view of representation and function the ITD activity map is a model of the MSO whereas the ILD activity map is a model of the LSO. The ILD and ITD likelihood maps can be seen as a model of the CNIC and the Bayesian Inference mechanism as a model for ENIC. The system is able to adapt to different robot heads just by learning some new reference maps in an echo-free room so that they are not dependent on the reverberation characteristics of any environment. The procedure generalizes across sounds and locations given the head model implicitly represented in the reference maps. The probabilistic approach for position estimation and propagation proofs to be a very efficient way to jointly analyze ITD and ILD and could also be extended to applications on tracking moving sound sources and/or to separate several speakers.

References

[1] K. Voutsas, G. Langner, J. Adamy, and M. Ochse, "A brain-like neural network for periodicity analysis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 1, pp. 12–22, Feb. 2004.

[2] Y. E. Cohen and E. Knudsen, "Maps versus clusters: Different representations of auditory space in the midbrain and forebrain," *Trends Neurosci.*, vol. 22, no. 3, pp. 128–135, Mar. 1999.

[3] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1996.

[4] D. Oertel, R. Fay, and A. Popper, *Integrative Functions in the Mammalian Auditory Pathway (Springer Handbook of Auditory Research)*. Heidelberg, Germany: Springer-Verlag, 2002.

[5] C. Cheng and G. Wakefield, "Introduction to head related transfer functions (hrtfs): Representation of hrtfs in time, frequency and space," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 231–249, Apr. 2001.

[6] J. Middlebrooks and D. Green, "Sound localization by human listeners," *Annu. Rev. Psychol.*, vol. 42, no. 1, pp. 135–159, 1991.

[7] E. Shaw, *The External Ear (Handbook of Sensory Physiology V/1: Auditory System, Anatomy Physiology)*. New York: Springer-Verlag, 1974.

[8] J. Adamy, K. Voutsas, and V. Willert, "A binaural sound localization system for mobile robots in low-reflecting environments," *Automatisierungstechnik*, vol. 51, no. 9, pp. 387–395, 2003.

[9] J. Rayleigh, "On our perception of sound direction," *Philos. Mag.*, vol. 13, no. 74, pp. 214–232, 1907.

[10] E. Jaynes, *Probability Theory—The Logic of Science*. Cambridge, MA: Cambridge Univ. Press, 2004.

[11] G. Langner, M. Sams, P. Heil, and H. Schulze, "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: Evidence from magnetoencephalography," *J. Comput. Physiol.*, vol. 181, no. 6, pp. 665–676, Dec. 1997.

[12] G. Ehret and R. Romand, *The Central Auditory System*. London, U.K.: Oxford Univ. Press, 1997.

[13] D. C. Fitzpatrick, S. Kuwada, and R. Batra, "Transformations in processing interaural time differences between the superior olivary complex and inferior colliculus: Beyond the Jeffress model," *Hear. Res.*, vol. 168, no. 1/2, pp. 79–89, Jun. 2002.

[14] S. Reyes. (2004). *The Auditory Central Nervous System*. [Online]. Available: http://serous.med.buffalo.edu/hearing/index.html

[15] L. O. Douglas, "Ascending efferent projections of the superior olivary complex," *Microsc. Res. Tech.*, vol. 51, no. 4, pp. 340–348, Nov. 2001.

[16] L. Jeffress, "A place theory of sound localization," *J. Comput. Physiol. Psychol.*, vol. 41, no. 1, pp. 35–39, 1948.

[17] S. Shamma, N. Shen, and P. Gopalaswamy, "Stereausis: Binaural processing without neural delay," *J. Acoust. Soc. Amer.*, vol. 86, no. 3, pp. 989–1006, Sep. 1989.

[18] S. Shamma, "On the role of space and time in auditory processing," *Trends Cogn. Sci.*, vol. 5, no. 8, pp. 340–348, Aug. 2001.

[19] B. Tran, T. Tran, "A sound localization system using Lyon's cochlear model and Lindemann's cross-correlation model," Dept. Elect. Eng., San Jose State Univ., San Jose, CA, Tech. Rep. No. 5, 1993.

[20] N. Bhadkamkar and B. Fowler, "A sound localization system based on biological analogy," in *Proc. IEEE Int. Conf. Neural Netw.*, San Francisco, CA, 1993, vol. 3, pp. 1902–1907.

[21] D. Rosen, D. Rumelhart, and E. Knudsen, "A connectionist model of the owl's sound localization system," in *Advances in Neural Information Processing Systems*, vol. 4. San Mateo, CA: Morgan Kaufmann, 1994, pp. 606–613.

[22] A. Handzel, S. Andersson, M. Gebremichael, and P. Krishnaprasad, "A biomimetic apparatus for sound source localization," in *Proc. 42nd IEEE Conf. Decision Control*, Maui, HI, 2003, pp. 5879–5885.

[23] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "Spiral vos final report: Part a, the auditory filterbank," Cambridge Electronic Design, Cambridge, U.K., Contract Rep. (Apu 2341), 1988. Internal Report.

[24] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched noise data," *Hear. Res.*, vol. 47, no. 1/2, pp. 103–138, Aug. 1990.

[25] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer Inc., Cupertino, CA, Tech. Rep. 35, 1993.

[26] F. Gustafsson, "Determining the initial states in forward-backward filtering," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 988–992, Apr. 1996.

[27] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*. London, U.K.: Prentice-Hall, 1989.

[28] J. Eggert, V. Willert, and E. Körner, "Building a motion resolution pyramid by combining velocity distribution," in *Pattern Recognition*, vol. 3175, Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 2004.

[29] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. Chichester, U.K.: Wiley, 2001.

[30] J. Bernardo and A. Smith, *Bayesian Theory (Wiley Series in Probability and Statistics)*. Chichester, U.K.: Wiley, 2004.

[31] J. Damaschke, "Towards a neurophysiological correlate of the precedence effect: From psychoacoustics to electroencephalography," Ph.D. dissertation, Medical Physics Dept., Univ. Oldenburg, Oldenburg, Germany, 2004.

[32] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[33] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robot. Auton. Syst.*, vol. 27, no. 4, pp. 199–209, Jun. 1999.

Volker Willert received the Dipl.Ing. degree in electrical engineering from the Darmstadt University of Technology, Darmstadt, Germany, in 2002.

From 2002 to 2005, he worked as a Ph.D. Student at the Control Theory and Robotics Laboratory, Institute of Automatic Control, Darmstadt University of Technology. Since 2005, he has been with the Honda Research Institute, Offenbach, Germany. His interests include image sequence processing and probabilistic modeling of cognitive biologically inspired systems.

Julian Eggert received the Ph.D. degree in physics from the Technical University of Munich, Munich, Germany, in 2000, where he was working at the Theoretical Biophysics Department.

He is currently with the Honda Research Institute, Offenbach, Germany. His interests comprise the dynamics of spiking neurons and neuronal assemblies, large-scale models for vision system, and gating in hierarchical neural networks via feedback and attention.

Jürgen Adamy received the Dipl.Ing. degree in electrical engineering and the Ph.D. degree in control theory from the University of Dortmund, Dortmund, Germany, in 1987 and 1991, respectively.

From 1991 to 1998, he worked as a Research Engineer and later as a Research Manager at the Siemens Research Center, Erlangen, Germany. Since 1998, he has been a Professor at the Darmstadt University of Technology, Darmstadt, Germany, and Head of its Control Theory and Robotics Laboratory.

Raphael Stahl received the Dipl.Ing. degree in electrical engineering from the Darmstadt University of Technology, Darmstadt, Germany, in 2004.

Since 2005, he has been the Chair of Automotive Engineering, Darmstadt University of Technology. His interests include auditory signal processing and graphical user interface programming for sound recording and analysis.

Edgar Körner received the Dr.Ing. degree in biomedical engineering and the Dr.Sci. degree in biocybernetics from the Technical University of Ilmenau, Ilmenau, Germany, in 1977 and 1984, respectively.

In 1988, he became Full Professor and Head of the Department of Neurocomputing and Cognitive Science, Technical University of Ilmenau. From 1992 to 1997, he was a Chief Scientist at Honda R&D Co., Wako, Japan. In 1997, he moved to Honda R&D Europe in Germany to establish the Future Technology Research Division, and since 2003, he has served as the President of the Honda Research Institute Europe GmbH, Offenbach, Germany. His research focus is on brainlike artificial neural systems for image understanding, smooth transition between signal-symbol processing, and self-organization of knowledge representation.