Binaural Sound Source Localisation and Tracking using a Dynamic Spherical Head Model

Christopher Schymura 1, Fiete Winter 2, Dorothea Kolossa 1, Sascha Spors 2

1 Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
2 Institute of Communications Engineering, University of Rostock, Germany

{christopher.schymura, dorothea.kolossa}@rub.de
{fiete.winter, sascha.spors}@uni-rostock.de

Abstract

This paper introduces a binaural model for the localisation and tracking of a moving sound source’s azimuth in the horizontal plane. The model uses a nonlinear state space representation of the sound source dynamics including the current position of the listener’s head. The state is estimated via an unscented Kalman Filter by comparing the interaural level and time differences of the binaural signal with semi-analytically derived localisation cues from a spherical head model. The localisation performance of the model is evaluated in combination with two different head movement approaches based on open- and closed-loop control strategies. The results show that adaptive strategies outperform non-adaptive ones and are able to compensate systematic deviations between the spherical head model and human heads.

Index Terms: Binaural localisation, head movements, machine hearing

1. Introduction

The human auditory system exploits the acoustic properties of the outer ear including torso, head and pinna for localising sound sources. Modelling the shape of the human head as a rigid sphere is a simple geometric approach to approximate the influence of the outer ear on the Head Related Transfer Functions (HRTFs). It has been shown in the past [1], that this approach provides insights in to how far the human head contributes to the localisation capabilities of the auditory system. Brungart and Rabinowitz [2] showed that the dependency of measured Interaural Level Differences (ILDs) and Interaural Time Differences (ITDs) on the distance of nearby sound sources can be reasonably approximated with a spherical head model. While the influence of the head shape is well covered, the effects of the pinna are however completely ignored by the model. This leads to significant deviations from measured HRTFs at high frequencies [3, p.100]. The model provides an analytic connection between the position of the sound source and the ILD and ITD. Contrary to other approaches [4, 5], no prior supervised training on measured HRTF datasets is necessary to establish this connection.

The model proposed in this study uses an Unscented Kalman Filter (UKF) [6] to infer the position of the sound source from measured ITDs and ILDs. Similar approaches have already been introduced in previous works, either using Kalman filtering techniques [7, 8] or particle filters [9]. Additionally, the effects of translatory movements and head rotations of the listener on localisation performance using a particle filter have been investigated in [10]. The results conform with the work of Wallach [11], indicating that rotational head movements improve azimuth localisation by resolving front-back ambiguities, which are likely to occur if sound sources are positioned within the cone of confusion [12]. A probabilistic framework with similar capabilities was introduced in [13], though only step-by-step head rotations were considered in this work. Extensions of the model [13] include the evaluation of different head-rotation strategies [14] and the increase of robustness in reverberant and noisy conditions [15].

This paper introduces a binaural model that is capable of conducting continuous head movements and analyses in how far the insufficiencies of the spherical head approximation can be compensated by considering acoustic scene dynamics and adaptive head movement strategies.

2. Binaural Model

The model introduced in this work is represented by a generic nonlinear dynamical system

\[ x_{k+1} = f(x_k, u_k) + v_k \]
\[ y_k = g(x_k) + w_k, \]

where \( x_k \) and \( y_k \) denote the hidden state and the observation vectors at time frame \( k \). The control input \( u_k \) is used to steer the head towards the desired orientation. \( f(\cdot) \) and \( g(\cdot) \) are nonlinear functions describing the model dynamics and the observations generated by the spherical head model, respectively. \( v_k \sim N(0, Q) \) and \( w_k \sim N(0, R) \) are zero-mean, Gaussian distributed noise vectors, with covariance matrices \( Q \) and \( R \). Throughout the course of this paper, it is assumed that both \( Q \) and \( R \) are either known in advance or can be estimated accordingly.

2.1. Model dynamics

The model dynamics (1) are represented by the 3-dimensional state vector

\[ x_k = [\phi_k, \dot{\phi}_k, \psi_k]^T \]

including the source position \( \phi_k \), the angular source velocity \( \dot{\phi}_k \) and the head orientation \( \psi_k \) (see Fig. 1). The process equations of the former two can be described by

\[ \phi_{k+1} = \phi_k + T\dot{\phi}_k + v_{\phi,k}, \quad v_{\phi,k} \sim N(0, \sigma_{\phi}^2) \]
\[ \dot{\phi}_{k+1} = \dot{\phi}_k + v_{\dot{\phi},k}, \quad v_{\dot{\phi},k} \sim N(0, \sigma_{\dot{\phi}}^2) \]

where \( T \) denotes the frame length of the Kalman filter in seconds. \( \sigma_{\phi}^2 \) and \( \sigma_{\dot{\phi}}^2 \) are the variances of the noise terms. The
In this work, the influence of continuous head rotations on the localisation performance is investigated. Two alternative strategies are proposed, relying on a purely feedforward and an adaptive feedback paradigm.

2.2.1. Static head position

If the control input is set to $u_k = 0 \forall k$, the look direction of the head will remain in its initial position. The static head position will serve as a baseline for all conducted experiments described in the following section.

2.2.2. Periodic scanning

The Periodic Scanning (PS) strategy is based on a feedforward controller

$$ u_k = \sin \left( 2\pi k \frac{T}{T_p} \right), $$

where $T_p$ denotes the duration of one scan cycle in seconds. This triggers the head to perform a bidirectional periodic rotation around the initial look direction.

2.2.3. Smooth Posterior Mean

In addition to the previously introduced PS strategy, a second approach using a closed-loop feedback controller

$$ u_k = \left[ 1 - \frac{1}{1 + |\phi_k - \psi_k|} \right] \cdot \text{sgn}(\phi_k - \psi_k) $$

is introduced and investigated. This approach is called the Smooth Posterior Mean (SPM) strategy, because the controller (11) steers the head on a smooth trajectory towards the posterior mean of the source position $\phi_k$ for each update of the UKF.

2.3. Binaural Front-End

The ILD and the ITD are the two main auditory cues for localising sound sources in the horizontal plane. The auditory front-end proposed by May et al. [4] is used to estimate the ILDs and the ITDs of a discrete binaural signal $s = [s_L, s_R]^T$. The subscripts L and R denote the signals corresponding to the left and the right ear, respectively. Both ear signals are sampled with a rate of $f_s = 1/T_s = 44.1kHz$. Each ear signal is then decomposed into $M = 32$ auditory channels using a phase compensated gammatone filterbank. The channel center frequencies $f_c$ are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5kHz. Half-wave rectification is applied to each frequency channel in order to extract the envelope. The ILDs and ITDs are then estimated for each frequency channel independently using non-overlapping, rectangularly windowed time frames with a length of 2048 samples ($T \approx 46.4ms$). The output of the binaural front-end is a vector

$$ \mathbf{b}(s_{k,L}, s_{k,R}) = \begin{bmatrix} \tau_1(s_{k,L}, s_{k,R}) & \ldots & \tau_M(s_{k,L}, s_{k,R}) \\ \delta_1(s_{k,L}, s_{k,R}) & \ldots & \delta_M(s_{k,L}, s_{k,R}) \end{bmatrix}^T $$

(12)

containing the ITDs $\tau_i(\cdot)$ and the ILDs $\delta_i(\cdot)$ for each frequency channel $i$ estimated for the $k$-th time frame. A generic implementation of the auditory front end used in this study is publicly available at [16].

2.4. Spherical Head Model

The nonlinear mapping function $g(x_k)$ introduced in Eq. (2) establishes a semi-analytical connection between the state vec-

Figure 1: For this work a right-hand coordinate system is used. The parameter $\psi_{ear}$, measuring the angle between the ears’ position and the z-axis is omitted for convenience.

The process equation of the look direction is represented as

$$ \psi_{k+1} = \text{sat}\left( \psi_k + T \psi_{\text{max}} \text{sat}(u_k) \right) + \nu_{\psi, k}, \nu_{\psi, k} \sim \mathcal{N}(0, \sigma_\psi^2), $$

(6)

where $\psi_{\text{max}}$ is the maximum angular velocity for the head rotation in radians per second, which is assumed to be constant. In order to model physical constraints of the maximum head displacement and restricted control inputs, two saturation functions $\text{sat}(x) = \min(|x|, x_{\text{max}}) \cdot \text{sgn}(x)$ are introduced in Eq. (6). Hence, the sets of possible angular head positions and control input values that the system can handle are defined by

$$ \mathcal{H} = \left\{ \psi_k \in \mathbb{R}^1 \mid |\psi_k| \leq \psi_{\text{max}} \right\}, $$

$$ \mathcal{U} = \left\{ u_k \in \mathbb{R}^1 \mid |u_k| \leq u_{\text{max}} \right\}, $$

(7)

(8)

where $\psi_{\text{max}}$ is the maximum rotational angle and $u_{\text{max}}$ is the system input limit. Furthermore, it is necessary to restrict the possible initial values for the look direction of the head within the range $\psi_0 \in [-\psi_{\text{max}}, \psi_{\text{max}}]$ on the frontal hemisphere.

By assuming uncorrelated disturbances affecting the state variables, the corresponding process noise covariance matrix can be expressed by

$$ Q = \begin{pmatrix} \sigma_{\phi}^2 & 0 & 0 \\ 0 & \sigma_{\phi}^2 & 0 \\ 0 & 0 & \sigma_{\psi}^2 \end{pmatrix} $$

(9)

It is worth noting, that the absolute angular position of the source and the look direction of the head are circular variables that lie within the range $(-\pi, \pi]$. This would imply that discontinuities are present in the state space, which would degrade the estimation results when using a generic UKF. To circumvent this issue, both quantities are treated as $2\pi$-periodic variables that are unbounded in $\mathbb{R}$, which can be handled by the spherical head model that will be presented in Section 2.4.

2.2. Integration of head-movement strategies

The proposed system allows the integration of continuous head movements through the control input $u_k$. A positive control input of $u_k = u_{\text{max}}$ triggers a clockwise head rotation with maximum angular velocity. Similarly, negative values induce a counter-clockwise head rotation.
tor (3) and the expected ILDs and ITDs. This is achieved by using an analytically derived binaural impulse response \( r(x_k) \), which solely depends on the apparent sound source azimuth, as an input to the auditory front-end described in 2.3. This allows to express the nonlinear function in the measurement model as

\[
g(x_k) = b(r_L(x_k), r_R(x_k)),
\]

where \( b \) is given in Eq. (12). The full measurement model is derived by inserting Eq. (13) into the generic measurement equation (2). In this paper, the disturbances affecting the measurements are assumed to be uncorrelated. Hence, the corresponding covariance matrix is defined as

\[
R = \begin{pmatrix} \sigma^2 I_M & 0_M \\ 0_M & \sigma^2 I_M \end{pmatrix},
\]

where \( I_M \) is the \( M \times M \) identity matrix and \( 0_M \) is a matrix of the same dimension containing only zeros. The variances of the noise affecting the ITDs and ILDs are denoted as \( \sigma^2_L \) and \( \sigma^2_R \).

### 2.4.1. Computation of Binaural Impulse Responses

The spherical head model [2, 17] is defined by three parameters, namely the head radius \( a \) and the angle pair \( (\theta_{ear}, \phi_{ear}) \) describing the position of both ears on the sphere (see Fig. 1). The time-frequency spectrum of the impulse response for the left ear \( r_L(\phi_k) \) is given as its Fourier transform

\[
R_L(x, \omega) = \frac{c}{4\pi\omega^2} \sum_{\nu=0}^{\nu_{max}} \sum_{\lambda=0}^{\lambda_{max}} h_L(\frac{\pi}{a} \lambda) (2\nu + 1) L_\nu \left( \frac{c}{a} \right) \sin(\theta_{ear}) \cos(\phi_k - \psi_k - \phi_L),
\]

where \( \phi_L = -\phi_R = \phi_{ear} \). The impulse response for the right ear \( r_R(\phi_k) \) can be generated accordingly. The spherical Hankel function of second kind and \( \nu \)-th order [18, sec. 10.1.1] is denoted by \( h_L(\cdot) \), while \( L_\nu(\cdot) \) symbolizes the \( \nu \)-th-degree Legendre polynomial [18, sec. 8.6.18]. The sound source position is described in polar coordinates by the distance \( d \) and the current apparent azimuth angle \( (\phi_k - \psi_k) \). The speed of sound is denoted by \( c \). The impulse responses for both ears are generated for each time frame \( k \) and truncated to 2048 samples each.

### 2.4.2. Implementation details

Two practical aspects have to be considered for the implementation of the spherical head model: First, the series involved in (15) has to be truncated sensibly with respect to accuracy and computational efficiency. According to [19, (43)], the series can be truncated at

\[
N = \left\lceil \frac{\pi\epsilon}{2\omega} \right\rceil + \max \left( 0, \left\lceil \ln \left( \frac{0.67848}{\epsilon} \right) \right\rceil \right),
\]

where \( \lceil \cdot \rceil \) and \( \epsilon \) denote the ceiling operator and the upper bound of the truncation error. Secondy, Eq. (15) has to be evaluated at regular and irregular frequencies \( f \) in order to perform an inverse discrete Fourier transform, whose result is the discrete binaural impulse response \( r_{L,R}(\phi_k) \). However, numerical instabilities are likely to occur, when dividing the two spherical Hankel functions (especially for high orders). A numerically stable approach uses a cascade of first- and second-order Infinite Impulse Response (IIR) filters, whose accumulated impulse response coincides with \( r_{L,R}(\phi_k) \). The coefficients of these filters are derived from Eq. (15) using digital filter design methods. For a detailed description of this approach, the reader is referred to [20, 21].

### 3. Evaluation

#### 3.1. Evaluation scenarios

The proposed binaural model was evaluated in two single-source localisation scenarios. In the first scenario, a static sound source was positioned at five different target azimuth angles: \( 30^\circ, 60^\circ, 90^\circ, 120^\circ \) and \( 150^\circ \), covering the whole range of the cone of confusion [12]. Since the localisation task was not restricted to the frontal plane, the binaural model had to deal with potential front-back ambiguities. The second evaluation scenario contained a dynamic scene, where the initial source position was chosen identically to the first scenario. Additionally, the scene involved a counter-clockwise, uniform circular movement of the source by \( 180^\circ \) over the total simulation time. No additional prior knowledge was provided to the model in either scenario.

#### 3.2. Experimental setup

For the scenarios described in Sec. 3.1, the ear signals are generated by convolving the source signal with a suitable HRTF. Based on the apparent source azimuth \( (\phi_k - \psi_k) \) the HRTF is reselected for each signal frame \( k \). Scene dynamics are simulated by cross-fading the HRTFs of subsequent frames. The binaural simulation tool used for this purpose is publicly available [22]. For the experiments, parts of the anechoic HRTF data set released by Wierstorf et al [23] are used. The HRTFs are measured with a Knowles Electronics Manikin for Acoustic Research (KEMAR), where the sound source is situated on a circle of 3m radius in the horizontal plane with an azimuthal resolution of 1°.

The target sound was speech signals taken from the GRID corpus [24]. The corpus consists of short utterances spoken by 34 native English speakers (18 male and 16 female speakers). The evaluation set was comprised of 100 randomly selected utterances from 5 male and 5 female speakers, where 10 utterances were taken per speaker. Speech pauses were excluded from the evaluation according to the corresponding alignments provided with the corpus. The individual utterances
were replicated to match a total duration of 5 seconds of speech. All experiments were conducted with identical model parameters listed in Tab. 1. State estimation was performed with a generic UKF [6], using the publicly available EKF/UKF Toolbox [25]. In all conducted experiments, the initial state was set to $\mathbf{x}_0 = [0, 0, 0]^T$, ensuring that the model is not provided with any prior knowledge about the source position and velocity. Localisation performance was measured for each utterance using the circular root mean square error (RMSE)

$$\text{cRMSE} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \min_{l \in \mathbb{Z}} \left( \hat{\phi}_k - \phi_k + 2\pi l \right)^2}$$

where $\hat{\phi}_k$ denotes the estimated source position at frame $k$, $\phi_k$ is the corresponding ground truth and $K$ is the total number of frames in one utterance.

### 3.3. Results and discussion

The achieved localisation performance for both scenarios is shown in Fig. 2. Each head rotation strategy was evaluated with all 100 utterances for 5 different (initial) positions of the target source. The average circular RMSEs depicted in Fig. 2 were computed as the mean over all utterance-level circular RMSEs for each experiment.

Fig. 2a shows localisation performance for the static scenario, indicating that controlled head movements yield improvements over open-loop controlled and static head positions. The SPM strategy outperforms both the static case baseline and the PS approach in all investigated source positions. The improvements over open-loop controlled and static head positions.

![Figure 2: Static scenario.](image1)

![Figure 2: Dynamic scenario.](image2)

In this study, a binaural model for localisation and tracking of sound sources, including continuous head rotations, was introduced. Experimental results have shown that closed-loop controlled head movements yield significant improvements in localisation performance over a pre-determined open-loop control strategy and no head movements in static and dynamic scenes. In particular, the investigated closed-loop strategy was able to reduce the effect of systematic errors of the underlying spherical head assumption of the model, in comparison with measured HRTFs from a KEMAR dummy head.

A next step for further investigations is the analysis of errors introduced by the spherical head model in order to improve robustness for localisation in different acoustic environments. The integration of sound distance as an additional state parameter is a second possibility for further extensions of the model. Additionally, the assumption of uncorrelated disturbances affecting the underlying state space model used in this study does not hold for practical applications. Therefore, it is necessary to further include and evaluate suitable techniques for the estimation of the corresponding covariance matrices describing the process and measurement noise. Due to the fact that the model proposed in this study is able to perform continuous head movements, it can also serve as a testbed for comparisons with human localisation performance assessed in listening tests.

### 4. Conclusions and future work

In this study, a binaural model for localisation and tracking of sound sources, including continuous head rotations, was introduced. Experimental results have shown that closed-loop controlled head movements yield significant improvements in localisation performance over a pre-determined open-loop control strategy and no head movements in static and dynamic scenes. In particular, the investigated closed-loop strategy was able to reduce the effect of systematic errors of the underlying spherical head assumption of the model, in comparison with measured HRTFs from a KEMAR dummy head.

A next step for further investigations is the analysis of errors introduced by the spherical head model in order to improve robustness for localisation in different acoustic environments. The integration of sound distance as an additional state parameter is a second possibility for further extensions of the model. Additionally, the assumption of uncorrelated disturbances affecting the underlying state space model used in this study does not hold for practical applications. Therefore, it is necessary to further include and evaluate suitable techniques for the estimation of the corresponding covariance matrices describing the process and measurement noise. Due to the fact that the model proposed in this study is able to perform continuous head movements, it can also serve as a testbed for comparisons with human localisation performance assessed in listening tests.

### 5. Acknowledgements

This research has been supported by EU FET grant TWO!EARS, ICT-618075.
6. References


