A Model-Based 2-D Head-Tracking Method Using Microphones at the Ears

Tobias Kabzinski, Jérôme Biot, Peter Jax
Institute of Communication Systems (IKS), *RWTH Aachen University*, Aachen, Germany
{kabzinski,jax}@iks.rwth-aachen.de

***Abstract*—Head-tracking is essential to dynamically reproduce binaural signals. If a crosstalk cancellation system is used for reproduction via loudspeakers, the head-tracking can be conducted based on acoustic delay estimates. Instead of estimating the delays between anchor sources and microphones at the listener's ears, we propose to obtain these delay estimates by continuously measuring the impulse responses between the loudspeakers and the microphones. Based on the measured delays, a nonlinear least-squares problem is presented to jointly estimate position and orientation in the horizontal plane. This problem formulation incorporates a novel angle-dependent parametric delay model to increase tracking precision. We show that our presented parametric delay model is superior to the free field and Woodworth's delay model in terms of head-tracking precision. Assuming free field conditions despite the presence of the listener's head is demonstrated to causes a systematic error. Moreover, we validate the proposed head-tracking method in a preliminary dynamic evaluation in an adaptive crosstalk cancellation system.**

***Index Terms*—Crosstalk cancellation, head-tracking, nonlinear least-squares**

I. Introduction

For the immersive reproduction of binaural audio signals, the listener's position and orientation in space must be tracked. These head-tracking data can either be used to modify the binaural synthesis [1] or to adapt pre-rendered binaural signals to the current user orientation [2]. Reproducing binaural signals via loudspeakers usually requires a dynamic crosstalk cancellation system, which also necessitates head-tracking data to design suitable filters [3], [4]. Instead of relying on external tracking devices, audio signals can also be employed to facilitate tracking. This is attractive if microphones are used anyway to adapt the crosstalk cancellation filters, as in [5].

To track the user's motion based on audio signals, multiple approaches using microphones at the ears have been proposed. Firstly, in [6], [7], anchor sources aid to estimate the acoustic propagation delay between loudspeakers and microphones. In a subsequent step, the user's position and orientation is estimated from these delays. This position estimation technique assumes free field conditions for the sound propagation despite the presence of the user's head. Consequently, we observed systematic errors. Secondly, the approach presented in [4], which is based on comparing the interaural time difference (ITD) contained in the desired binaural playback signal and the ITD contained in the microphone signal, estimates the orientation around the vertical axis only. While lacking position estimates, the ITD error-based approach does not require potentially disturbing anchor signals.

In this paper, a head-tracking method is presented which not only avoids using potentially disturbing anchor sources to obtain propagation delay estimates but can also include a more sophisticated propagation delay model to mitigate the systematic error arising from the inadequate free field assumption. Our investigations focus on the application in crosstalk cancellation systems but are not limited to those. We suggest to estimate the impulse responses between the loudspeakers and the microphones at the ears online — a challenging system identification problem, as discussed in [5]. The impulse response estimates provide access to the propagation delays without requiring anchor sources. Based on a parametric propagation delay model, which can easily be personalized, position and orientation estimates are found as the solution of a nonlinear optimization problem.

The remainder of this paper is structured as follows: Sec. II describes the proposed head-tracking method and introduces the parametric propagation delay model. In Sec. III, the proposed delay model is compared to other delay models, and the head-tracking performance is evaluated in two experiments. Sec. IV concludes the paper.

II. Binaural Head-Tracking Method

The goal of acoustic, specifically binaural, 2-D head-tracking is to estimate the user's position $\mathbf{p}_\mathrm{U} = [x_\mathrm{U}, y_\mathrm{U}]^\mathrm{T}$ and orientation ϕ_U (rotation around the z-axis) exploiting the microphone signals at the user's ears. In contrast to the binaural source localization problem (e.g., [8]), the source signals are accessible here. It is assumed that the microphone positions are similar to those used for measuring head-related transfer functions (HRTFs) with open ear canals. According to [9], placing the microphones slightly outside the ear canal does not change the HRTFs fundamentally. Furthermore, the two loudspeaker positions, $\mathbf{p}_{\mathrm{S}_j} = \left[x_{\mathrm{S}_j}, y_{\mathrm{S}_j}\right]^\mathrm{T}, j = 1, 2$, are assumed to be known.

A. Review of Tikander's Method

Tikander et al. have proposed to obtain delay estimates $\hat{t}_{ij}$, between source j and receiver $i \in \{\mathrm{L}, \mathrm{R}\}$, by evaluating the cross-correlation of the anchor signal from loudspeaker j and the microphone signal at receiver i [7]. As they refer to standard methods to solve the underlying geometric problem, known as trilateration or "circulation" [7], we believe that sound propagation under free field conditions is assumed implicitly by converting the delays to distances.

To create a reference implementation of their method, our interpretation is as follows: Via trilateration, we estimate separately the two ear positions $\hat{\mathbf{p}}_{\mathrm{E}_i} = [\hat{x}_{\mathrm{E}_i}, \hat{y}_{\mathrm{E}_i}]^{\mathrm{T}}$. Then, the average serves as the estimated center of the user's head, i.e.,

$$\hat{\mathbf{p}}_{\mathrm{U}} = [\hat{x}_{\mathrm{U}}, \hat{y}_{\mathrm{U}}]^{\mathrm{T}} = (\hat{\mathbf{p}}_{\mathrm{E_L}} + \hat{\mathbf{p}}_{\mathrm{E_R}})/2. \quad (1)$$

If the trilateration yields two solutions, we choose the one such that the user is positioned in front of the loudspeakers.

In [7], it is suggested to use an ITD model to estimate the relative angles between the user orientation and each loudspeaker. We choose the Woodworth ITD model, e.g., [10],

$$\mathrm{ITD}(\phi) = \begin{cases} \frac{a}{c}(\phi + \sin(\phi)), & \text{for } 0 \le |\phi| \le \frac{\pi}{2} \\ \frac{a}{c}(\mathrm{sgn}(\phi)\pi - \phi + \sin(\phi)), & \text{for } \frac{\pi}{2} < |\phi| \le \pi \end{cases}.$$

Here, a denotes the head radius, c is the speed of sound and $\mathrm{sgn}(\cdot)$ symbolizes the sign function. For positive ϕ, the look direction is left of the source. Due to the brevity in [7], it remains unclear how to obtain one user orientation angle from multiple (possibly contradictory) relative angles to the different loudspeakers. Thus, to obtain a single estimate of ϕ_{U}, we suggest to solve the nonlinear least-squares problem

$$\min_{\phi_{\mathrm{U}}} \sum_{j=1,2} \left(\hat{t}_{\mathrm{L}j} - \hat{t}_{\mathrm{R}j} - \mathrm{ITD}(\phi_{\mathrm{U}} - \Delta\phi_j)\right)^2. \quad (2)$$

This aims at finding the orientation ϕ_{U} which best explains *all* the observed ITDs. The relative angle between source j and the estimated user position can be calculated as $\Delta\phi_j = \mathrm{atan2}\left(y_{\mathrm{S}_j} - \hat{y}_{\mathrm{U}}, x_{\mathrm{S}_j} - \hat{x}_{\mathrm{U}}\right)$, where atan2 is the four-quadrant extension of the inverse tangent function. From the estimated ear positions (1), an orientation estimate, which can be used as an initial value for solving (2), can be derived as

$$\hat{\phi}_{\mathrm{U}} = \mathrm{atan2}(\hat{y}_{\mathrm{E_R}} - \hat{y}_{\mathrm{E_L}}, \hat{x}_{\mathrm{E_R}} - \hat{x}_{\mathrm{E_L}}). \quad (3)$$

B. Proposed Problem Formulation

We propose a binaural 2-D head-tracking algorithm which is different from the approaches in the literature w.r.t. three aspects. Firstly, we measure the impulse responses $\hat{h}_{ij}(k)$ between the loudspeakers and the microphones only from the desired playback signals, as in [5]. This way, no potentially disturbing anchor signals are required. Secondly, we propose to *jointly* estimate position and orientation based on delays extracted from the measured impulse responses. Thirdly, to avoid systematic errors, we encourage using a propagation delay model different from the free field model.

Similarly to [11], estimates of position and orientation, $\hat{\mathbf{p}}$ and $\hat{\phi}_{\mathrm{U}}$, respectively, are found as those values which best explain the observed delays for a given delay model. This yields the nonlinear least-squares optimization problem

$$\min_{\mathbf{p}_{\mathrm{U}}, \phi_{\mathrm{U}}} \sum_{i \in \{\mathrm{L,R}\}} \sum_{j=1,2} \left(\hat{t}_{ij} - f_{i,\mathcal{M}}(\mathbf{p}_{\mathrm{S}_j}, \mathbf{p}_{\mathrm{U}}, \phi_{\mathrm{U}} | \boldsymbol{\theta}_{\mathcal{M}})\right)^2, \quad (4)$$

for which (1) and (3) provide initial values for the solver. The propagation delay model $f_{i,\mathcal{M}}$, with parameter set $\boldsymbol{\theta}_{\mathcal{M}}$, consists of a distance-dependent contribution and an angle-dependent contribution, i.e., $f_{i,\mathcal{M}}(\mathbf{p}_{\mathrm{S}_j}, \mathbf{p}_{\mathrm{U}}, \phi_{\mathrm{U}}) = \tau_0(\Delta\mathbf{p}_j) + \tau_{\mathcal{M},i}(\Delta\phi_j)$. The position difference between loudspeaker j and the user's head center determines the distance-dependent delay: $\tau_0(\Delta\mathbf{p}_j) = \|\Delta\mathbf{p}_j\|/c$ with $\Delta\mathbf{p}_j = \|\mathbf{p}_{\mathrm{S}_j} - \mathbf{p}_{\mathrm{U}}\|$ and $\|\cdot\|$ denoting the Euclidean norm. The relative angle between the source j and the user orientation is $\Delta\phi_j = \mathrm{atan2}\left(y_{\mathrm{S}_j} - y_{\mathrm{U}}, x_{\mathrm{S}_j} - x_{\mathrm{U}}\right)$ and determines the angle-dependent delay offset w.r.t. the head center.

For completeness, we state the different delay models which are compared later. From simple geometric considerations, the angle-dependent free field model with antipodal ears yields

$$\tau_{\mathcal{F},i}(\phi) = -(a/c)\cos(\phi + \phi_{\mathrm{E}_i}) \quad (5)$$

with $\phi_{\mathrm{E_L}} = +\pi/2$ and $\phi_{\mathrm{E_R}} = -\pi/2$. The (modified) Woodworth delay model, as in [12], is given by

$$\tau_{\mathcal{W},i}(\phi) = \begin{cases} -\frac{a}{c}\cos(\phi + \phi_{\mathrm{E}_i}), & \text{for } 0 \le |\phi + \phi_{\mathrm{E}_i}| \le \frac{\pi}{2} \\ s\frac{a}{c}\left||\phi + \phi_{\mathrm{E}_i}| - \frac{\pi}{2}\right|, & \text{for } \frac{\pi}{2} \le |\phi + \phi_{\mathrm{E}_i}| < \pi \end{cases}. \quad (6)$$

For $s = 1$, the classical Woodworth model is obtained, which we found to be improved by the modification $s \neq 1$ (cf. Sec. III-A). Alternatively, our proposed parametric delay model, which is described next, can be employed.

C. Parametric Delay Model

The physically motivated Woodworth model in (6) has served as a delay model in the past, especially as a foundation of HRTF models and, more specifically, ITD models [10], [12], [13]. For the application to binaural head-tracking, however, the model has three limitations. Firstly, there is only a single free parameter, the head radius a. While simple, the modeling quality remains limited. Secondly, the model reflects the true delays only for high frequencies, i.e., if the wavelength is small compared to the size of the head. Thirdly, extracting this exact delay from a real HRTF is difficult due to the complexity of HRTFs, especially due to the presence of the pinna (cf. [13]). A reliable delay estimation method, e.g., detecting when the impulse response envelope reaches for the first time a certain value relative to its maximum value, does not provide the exact delays predicted by the Woodworth model. Yet, a generic model should be able to incorporate this dependency on the delay estimation method. To be clear: If the same delay estimation method is used in training the model and in operation, resulting positions and orientation angles remain unbiased, even though the delay estimates themselves might be biased.

Having analyzed the angle-dependent delay contributions based on the variety of head shapes, ear positions, and other anthropometric characteristics present in the HRTF database [14], we propose a parametric delay model (PDM). For our analysis, we chose the above envelope-based threshold method with a threshold of $-10\,\mathrm{dB}$ and upsampling by factor 50. The PDM is similar to Woodworth's model in structure, featuring line segments, sine-like segments and connecting polynomial segments. Furthermore, we include multiple angle-delay pairs to increase adaptability and to be able to capture the dependency on the delay estimation method itself. As opposed to the modified Woodworth model in (6) with $s \neq 1$, the parametric model is constructed such that the derivative can be continuous except at ϕ_1. The proposed model is shown in Fig. 1 for the left receiver and is specified in a piece-wise manner:

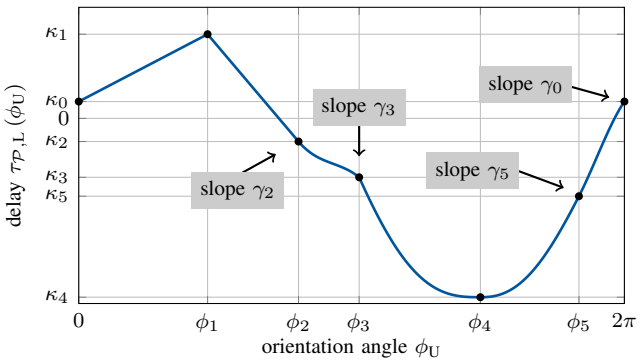

Fig. 1. Parametric model for angle-dependent propagation delay to left ear.

$$\tau_{\mathcal{P},\mathrm{L}}(\phi) = \begin{cases} \ell(\phi|\kappa_0,\kappa_1,0,\phi_1), & \text{for } 0 \le \phi \le \phi_1 \\ \ell(\phi|\kappa_1,\kappa_2,\phi_1,\phi_2), & \text{for } \phi_1 \le \phi \le \phi_2 \\ \wp(\phi|\kappa_2,\kappa_3,\phi_2,\phi_3,\gamma_2,\gamma_3), & \text{for } \phi_2 \le \phi \le \phi_3 \\ \alpha_{\text{down}}(\phi|\kappa_3,\kappa_4,\phi_3,\phi_4,\gamma_3), & \text{for } \phi_3 \le \phi \le \phi_4 \\ \alpha_{\text{up}}(\phi|\kappa_4,\kappa_5,\phi_4,\phi_5,\gamma_5), & \text{for } \phi_4 \le \phi \le \phi_5 \\ \wp(\phi|\kappa_5,\kappa_0,\phi_5,2\pi,\gamma_5,\gamma_0), & \text{for } \phi_5 \le \phi < 2\pi \end{cases}. \quad (7)$$

Four types of piece-wise functions changing from value κ_b to κ_e between ϕ_b and ϕ_e, are defined, namely, a line segment $\ell(\phi|\kappa_\mathrm{b},\kappa_\mathrm{e},\phi_\mathrm{b},\phi_\mathrm{e}) = \kappa_\mathrm{b} + (\kappa_\mathrm{e}-\kappa_\mathrm{b})\frac{(\phi-\phi_\mathrm{b})}{(\phi_\mathrm{e}-\phi_\mathrm{b})}$, a sine-like downward arc with slope γ_b at the left boundary

$$\alpha_{\text{down}}(\phi|\kappa_\mathrm{b},\kappa_\mathrm{e},\phi_\mathrm{b},\phi_\mathrm{e},\gamma_\mathrm{b}) = \kappa_\mathrm{e} + (\kappa_\mathrm{b}-\kappa_\mathrm{e})\left(1+\sin\left(\frac{\phi-\phi_\mathrm{b}}{\phi_\mathrm{e}-\phi_\mathrm{b}}\frac{\pi}{2}+\pi\right)\right)^{\frac{2\gamma_\mathrm{b}}{\pi}\frac{\phi_\mathrm{e}-\phi_\mathrm{b}}{\kappa_\mathrm{e}-\kappa_\mathrm{b}}},$$

a sine-like upward arc with slope γ_e at the right boundary

$$\alpha_{\text{up}}(\phi|\kappa_\mathrm{b},\kappa_\mathrm{e},\phi_\mathrm{b},\phi_\mathrm{e},\gamma_\mathrm{e}) = \kappa_\mathrm{b} + (\kappa_\mathrm{e}-\kappa_\mathrm{b})\left(1+\sin\left(\frac{\phi-\phi_\mathrm{b}}{\phi_\mathrm{e}-\phi_\mathrm{b}}\frac{\pi}{2}-\frac{\pi}{2}\right)\right)^{\frac{2\gamma_\mathrm{e}}{\pi}\frac{\phi_\mathrm{e}-\phi_\mathrm{b}}{\kappa_\mathrm{e}-\kappa_\mathrm{b}}},$$

and a third-order polynomial with given slopes γ_b and γ_e at the boundaries $\wp(\phi|\kappa_\mathrm{b},\kappa_\mathrm{e},\phi_\mathrm{b},\phi_\mathrm{e},\gamma_\mathrm{b},\gamma_\mathrm{e}) = \sum_{n=0}^{3} p_n \phi^n$. Its coefficients p_n can be found as a result of Hermite interpolation or by solving a system of linear equations.

To achieve a continuous first derivative of $\tau_{\mathcal{P},\mathrm{L}}(\phi)$, the slopes at the interval boundaries must be chosen to be $\gamma_0 = \frac{\kappa_1-\kappa_0}{\phi_1}$ and $\gamma_2 = \frac{\kappa_2-\kappa_1}{\phi_2-\phi_1}$. For the right ear, a mirrored model, with its own parameter set, can be set up replacing ϕ by $2\pi-\phi$.

D. Optimization of Model Parameters

To minimize the model error, it is suggested to optimize the model parameters on a given training set $\mathcal{T}$ of HRTF sets from one or more subjects. The model parameters are given by the sets $\boldsymbol{\theta}_\mathcal{F} = \{a\}$, $\boldsymbol{\theta}_\mathcal{W} = \{a, s\}$ and $\boldsymbol{\theta}_\mathcal{P} = \left\{\kappa_0^{(i)},\ldots,\kappa_5^{(i)},\phi_1^{(i)},\ldots,\phi_5^{(i)},\gamma_3^{(i)},\gamma_5^{(i)} \,|\, i=\mathrm{L},\mathrm{R}\right\}$. Each subject $m \in \mathcal{T}$ is assumed to be at $\mathbf{p}_\mathrm{U} = [0,0]^\mathrm{T}$ and $\phi_\mathrm{U} = 0$. Then, the single source takes those positions $\mathbf{p}_{\mathrm{S}_1}^{(n)}, n = 1,\ldots,N$, such that the user positions and orientations match those from the HRTF measurement in the horizontal plane. This way, an HRTF interpolation, which might introduce an interpolation error, is not required. An optimized set of model parameters $\boldsymbol{\theta}_\mathcal{M}$ can be found solving

$$\min_{\boldsymbol{\theta}_\mathcal{M}} \sqrt{\sum_{\substack{i\in\{\mathrm{L},\mathrm{R}\},\\ m\in\mathcal{T},\\ n=1,\ldots,N}} \left(t_{i1}^{(m,n)} - f_{i,\mathcal{M}}\left(\mathbf{p}_{\mathrm{S}_1}^{(n)}, \mathbf{p}_\mathrm{U}, \phi_\mathrm{U}|\boldsymbol{\theta}_\mathcal{M}\right)\right)^2}, \quad (8)$$

where $t_{i1}^{(m,n)}$ denotes the delay extracted from the given head-related impulse response (HRIR) $h_{i1}^{(m,n)}(k)$ of subject m for source 1 at position n.

For a fair comparison, the parameters of all delay models must be optimized. Using the Quasi-Newton method to solve (8) for the free field model or for the Woodworth model yields good results. However, the solution for the high-dimensional parametric delay model obtained in this way often appears to be only a local minimum, with obvious large errors. Therefore, we suggest to apply an iterative optimization approach, in which only one variable is optimized at a time and the others are kept at fixed values. This process is repeated until each parameter has been optimized $\mathcal{I}$ times.

III. Experimental Evaluation

To evaluate the parametric delay model and the proposed head-tracking method, we conducted three experiments.

A. Comparison of Delay Models

To compare the quality of the models, their parameters are optimized as described in Sec. II-D, and $\mathcal{I} = 5$ iterations are conducted. Each optimization step is conducted using MATLAB's Quasi-Newton method implementation in `fminunc`. The initial values for the free field and Woodworth's model are $a = 8.75\,\mathrm{cm}$ and $s = 1$, while the parametric delay model is initialized such that it yields values very close to those of the Woodworth model. The speed of sound is assumed to be $c = 343\,\mathrm{m/s}$ and the sample rate is $12\,\mathrm{kHz}$. The delays $\hat{t}_{ij}$ are estimated as described in Sec. II-C. We found that this delay estimation method is quite robust, especially for contralateral HRIRs. As a dataset, we use the numerically simulated HRTFs of 93 unique human subjects from [14]. The corresponding indices m shall be in $\mathcal{T}_{\text{HUTUBS}}$. The parameter optimization on the 72 measurement positions in the horizontal plane for each of the 93 subjects is repeated twice: personally for each subject m_0 , i.e., $\mathcal{T} = \{m_0\}$, and in a leave-one-out fashion, i.e., $\mathcal{T} = \mathcal{T}_{\text{HUTUBS}} \setminus \{m_0\}$. To quantify the model error, the residual error

$$\varepsilon_{i1}^{(m_0,n)} = \left| \hat{t}_{i1}^{(m_0,n)} - f_{i,\mathcal{M}}\left(\mathbf{p}_{\mathrm{S}_1}^{(n)}, \mathbf{p}_\mathrm{U}, \phi_\mathrm{U}|\boldsymbol{\theta}_\mathcal{M}\right)\right|, \quad (9)$$

for subject m_0 and position n, is calculated for $N = 360$ uniformly spaced test orientations ϕ_U in the horizontal plane. To simulate these orientations, the spherical spline interpolation method [15], as implemented in [16], is used[1].

Fig. 2 visualizes the delay-angle dependency of the personally optimized delay models for subject $m_0 = 3$. It can be seen that the PDM captures the measured delay behavior most

[1]To reduce the computational demand, only six neighbors are used.

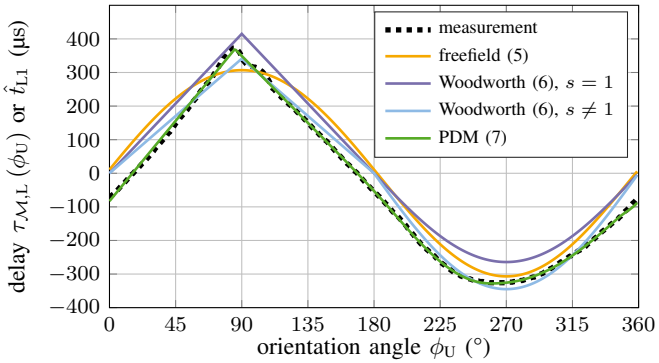

Fig. 2. Exemplary delay models personally optimized for subject 3.

TABLE I
COMPARISON OF ERRORS FOR DIFFERENT DELAY MODELS

	personal		leave-one-out	
	mean error (µs)	std error (µs)	mean error (µs)	std error (µs)
freefield (5)	51.1	28.0	51.4	28.8
Woodworth (6), $s = 1$	57.6	19.2	57.9	20.5
Woodworth (6), $s \neq 1$	27.8	19.8	29.7	20.8
PDM (7)	4.2	6.2	14.0	12.8

precisely. Table I shows the means and standard deviations of the $2{\cdot}93{\cdot}360 = 66960$ errors (9) in the eight cases. The fact that the mean error does not decrease much when changing from "leave-one-out" to "personal" indicates that the models other than the PDM cannot represent the delay behavior properly. The modified Woodworth model roughly cuts the mean error in half. Yet, the proposed PDM by far outperforms the other models in both the personal and the leave-one-out case. This implies good suitability of the PDM for head-tracking.

B. Head-Tracking Performance in Static Conditions

To showcase the systematic error in case of model mismatch, an experiment with the following setup is conducted. On the one hand, the acoustic transfer functions represent free field conditions (FF), simulated as fractional delay filters. On the other hand, the presence of a rigid sphere (RS) is simulated, as described in [17], as an approximation for the presence of the listener's head. The head radius is $a = 8.75\,\mathrm{cm}$. The loudspeakers, positioned at angles $\pm 45°$, are $2.5\,\mathrm{m}$ away from the origin. The user position is $\mathbf{p}_\mathrm{U} = [0, 0]^\mathrm{T}$. The delays $\hat{t}_{ij}$ are estimated as the time when the envelope of the 50-times upsampled impulse response $h_{ij}(k)$ reaches its maximum. This delay estimation method yields the exact values for the free field model, but it is not reliable for real HRTFs.

Fig. 3 shows the position and orientation estimation errors for different user orientations. Due to symmetry, the analysis is restricted to rotations of up to 180°. If system and model match, the resulting errors are very close to zero, i.e., either the trilateration method (1) is applied to the free field impulse responses or (4) is used in combination with the free field model (5). If the Woodworth model (6) is applied to the rigid sphere transfer functions, the errors are still small. In this case, the delay estimates provided by the delay estimation method do not exactly match the values of the Woodworth model and

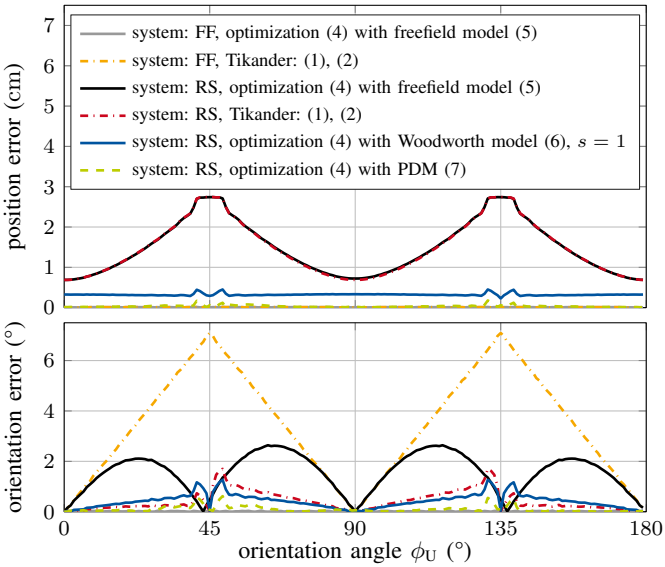

Fig. 3. Comparison of estimation errors under ideal static conditions.

small angular errors occur. The PDM, optimized on the rigid sphere's delays estimated with same maximum envelope delay estimation method, yields lower position and angular errors, as expected. If free field conditions are assumed in the model but the actual system is a rigid sphere, a systematic error can be seen. In this example, the errors reach about $2.7\,\mathrm{cm}$ in position and $2.5°$ in angle. When one of the "ears" on the sphere is contralateral, corresponding to an orientation of $45°$ or $135°$, the errors in position are especially strong and can be explained as follows. Due to the longer path around the sphere, the observed delays are larger than under free field conditions. Hence, these longer delays can be best explained by a slight position offset in the trilateration method and in the proposed method if free field conditions are assumed. In a nutshell, this experiment illustrates that a precise delay model is crucial to achieve low head-tracking estimation errors.

C. Head-Tracking Performance in Dynamic Conditions

To demonstrate the feasibility of the proposed head-tracking method, an adaptive crosstalk cancellation setup, as in [5], with the geometric setup as in Sec. III-B is considered. A continuous head rotation of angular velocity of $5°/\mathrm{s}$ is simulated in an anechoic environment. To achieve this, the simulated HRTFs of subject 3 from [14] are interpolated every $10\,\mathrm{ms}$ using the spherical spline interpolation, as described in Sec. III-A. The binaural playback signal is a $10\,\mathrm{s}$-long repeated excerpt from a binaural orchestra recording, and the crosstalk cancellation filters are updated every $10\,\mathrm{ms}$ with the frequency-domain least-squares method [18]. To obtain impulse response estimates, we resort to a multichannel extension of the time-domain Kalman filter [19] in combination with the measurement noise estimation method from [20]. This multichannel system identification will certainly be the limiting factor for tracking fast movements, but the system identification problem, especially for moving subjects, must be considered a research topic of its own. Using a relatively low sampling rate of $12\,\mathrm{kHz}$ allows to obtain sufficiently precise delay estimates,

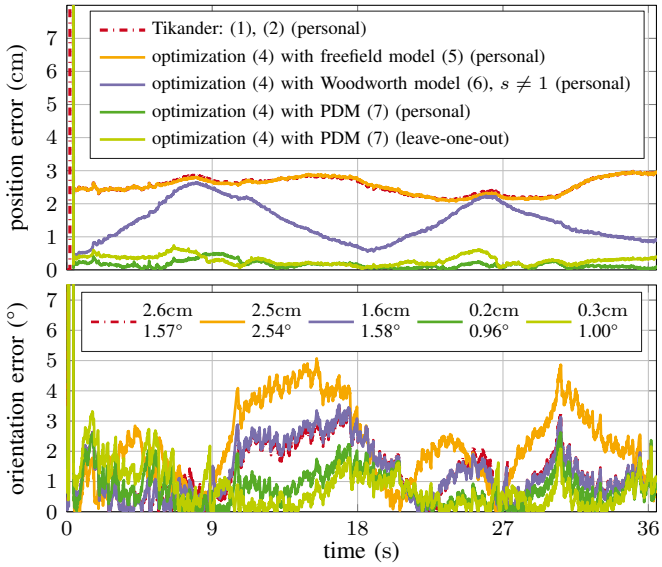

Fig. 4. Comparison of estimation errors in dynamic conditions. The lower legend shows the RMS errors in position and orientation (for times after 0.5 s).

with the delay estimation method from Sec. II-C, at a simulated SNR of 30 dB at the microphones. Furthermore, we only estimate 84 filter coefficients per channel and assume 48 zero-valued coefficients before the nonzero coefficients to avoid estimating the propagation delay part of the impulse responses. This would unnecessarily slow down the identification of the impulse responses. In the real-world, the audio interface latency would need to be considered here, too.

Fig. 4 visualizes the estimation errors in position and orientation for different estimation methods. The large errors before 0.5 s result from initially bad impulse response estimates. While both free field-based methods yield a relatively constant position error of about 2 cm to 3 cm, using the Woodworth model in the proposed method reduces this error. However, the peak errors remain similarly high when one of the ears is contralateral (around 9 s and 27 s). As the proposed PDM fits the measured delays much better (cf. Sec. III-A), a position estimation with higher precision is achieved and the systematic position error, as in Fig. 3, vanishes. Surprisingly, the PDM with leave-one-out optimization performs similarly to the personalized PDM. The RMS angular errors are comparable for all methods and similar to the values reported in [21].

IV. Conclusion

In this paper, a head-tracking method using microphones at the listener's ears has been proposed. Instead of using anchor signals, delay estimates are obtained from estimated impulse responses based on the desired playback signal. The limits imposed by current system identification methods, especially in reverberant environments, and their impact on the delay estimation error, need further investigation. In the proposed method, user position and orientation are found as the solution of an optimization problem, which includes a novel angle-dependent parametric delay model. With this model, both with and without personalization, the position estimation errors are reduced when comparing with free field-based methods.

References

[1] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer Science & Business Media, 2007.

[2] S. Nagel and P. Jax, "Dynamic binaural cue adaptation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 96–100.

[3] T. Lentz, "Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments," *Journal of the Audio Engineering Society*, vol. 54, no. 4, pp. 283–294, 2006.

[4] Y. Lacouture-Parodi and E. A. Habets, "Crosstalk cancellation system using a head tracker based on interaural time differences," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, 2012.

[5] T. Kabzinski and P. Jax, "An adaptive crosstalk cancellation system using microphones at the ears," in *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[6] M. Karjalainen, M. Tikander, and A. Harma, "Head-tracking and subject positioning using binaural headset microphones and common modulation anchor sources," in *2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, 2004.

[7] M. Tikander, A. Harma, and M. Karjalainen, "Binaural positioning system for wearable augmented reality audio," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2003, pp. 153–156.

[8] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2009.

[9] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[10] N. L. Aaronson and W. M. Hartmann, "Testing, correcting, and extending the Woodworth model for interaural time difference," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 817–823, 2014.

[11] S. Nagel, T. Kabzinski, S. Kühl, C. Antweiler, and P. Jax, "Acoustic head-tracking for acquisition of head-related transfer functions with unconstrained subject movement," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[12] D. Romblom and H. Bahu, "A revision and objective evaluation of the 1-pole 1-zero spherical head shadowing filter," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[13] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, 1998.

[14] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3d head meshes, anthropometric features, and headphone impulse responses," *Journal of the Audio Engineering Society*, vol. 67, no. 9, pp. 705–718, 2019.

[15] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[16] F. Brinkmann and S. Weinzierl, "Aktools—an open software toolbox for signal acquisition, processing, and inspection in acoustics," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[17] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[18] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 189–194, 1998.

[19] S. Liebich, J. Fabry, P. Jax, and P. Vary, "Time-domain Kalman filter for active noise cancellation headphones," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 593–597.

[20] T. Strutz, "Estimation of measurement-noise variance for variable-step-size NLMS filters," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019.

[21] Y. Lacouture-Parodi and E. A. Habets, "Application of particle filtering to an interaural time difference based head tracker for crosstalk cancellation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 291–295.