Active Localization of Sound Sources with Binaural Models

Christopher Schymura, Juan Diego Rios Grajales, Dorothea Kolossa
Institute of Communication Acoustics, Ruhr-Universität Bochum, 44801 Bochum, Germany
Email: christopher.schymura@rub.de, juan.riosgrajales@rub.de, dorothea.kolossa@rub.de

Abstract

During past decades, a variety of different models for sound source localization has been developed in the context of binaural hearing. This paper introduces extensions to recently developed models, which consider localization as an active process, inspired by the ability of humans to conduct head movements. A machine hearing system is presented that utilizes different head motion strategies via a closed-loop feedback control scheme to localize sources at any angular position in the horizontal plane. The proposed framework is based on representing the localization problem as a dynamical system, where the source position is inferred from interaural time- and level-difference measurements by recursive Bayesian estimation techniques. This approach enables the system to be used for online processing on a frame-by-frame basis and to adapt accordingly to changes in a dynamic acoustic scene. A systematic evaluation of different head movement strategies shows that closed-loop feedback is able to substantially increase localization performance in different acoustic conditions.

Introduction

An essential aspect of auditory scene analysis (ASA) is the localization of sound sources in the environment surrounding the listener. The human auditory system is capable of precisely locating and separating different sound sources, even in noisy and reverberant environments. Machine hearing systems differ from human listeners in a number of important respects: First, machine hearing systems typically assume that the acoustic sensors are fixed in their position. In contrast, human hearing relies on motion to provide listeners with information about changes in interaural time differences (ITDs) and interaural level differences (ILDs) which can be used to disambiguate the location of a sound source [1]. Secondly, machine hearing systems typically follow a bottom-up paradigm, which also stands in contrast to auditory processing, in which top-down feedback is known to play an important role [2].

Recently, several machine hearing systems have been proposed which include top-down feedback via rotational head movements to improve sound source localization [3, 4, 5]. This study proposes an extension of the work introduced in [5], which tackles the problem of sound source localization from a control-theoretical standpoint. Specifically, the dynamics of the rotational movement of the acoustic sensors and potential dynamics of sound sources are modeled as a stochastic, nonlinear state space representation. This is used to infer the position of a sound source and the look direction of the head from measurement data with recursive Bayesian estimation techniques. The motion of the head is controlled via closed-loop feedback, taking into account the estimated states. In contrast to [5], this study proposes two novel feedback strategies for rotational head movement and assesses the localization performance of the system in simulated reverberant conditions.

Binaural front-end

A binaural front-end as proposed in [6] is used to extract monaural and binaural features from binaural ear signals, sampled with a rate of $f_\mathrm{s} = 44,1\,\mathrm{kHz}$. Each channel of the ear signals is decomposed into $L = 32$ auditory channels using a phase compensated gammatone filterbank. The filter center frequencies are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 8 kHz [7]. Half-wave rectification and low-pass filtering is applied to each frequency channel to model the effect of the inner hair cells (IHCs) [8]. Subsequently, monaural and binaural features are extracted using non-overlapping, rectangularly windowed time frames with a length of 20 ms.

The localization model used in this study is based on two primary binaural cues, namely ITDs and ILDs. The ITD $\hat{\tau}_{kl}$ between the left and the right ear signal is estimated for each time frame k and frequency channel l by locating the time lag that corresponds to the maximum of the interaural cross-correlation function. ILDs are estimated analogously by comparing the frame-based energy of the left and right ear IHC signals. They are denoted as $\hat{\delta}_{kl}$ and expressed in dB. Across-frequency integration is conducted for both binaural cues to compensate for outliers and estimation errors in specific frequency bands [REF], yielding a 2-dimensional binaural feature vector

$$\begin{bmatrix}\hat{\tau}_k \\ \hat{\delta}_k\end{bmatrix} = \begin{bmatrix}\sum_{l=1}^{L}\hat{\tau}_{kl} & \sum_{l=1}^{L}\hat{\delta}_{kl}\end{bmatrix}^T \tag{1}$$

at each time frame.

Localization model

The localization model proposed in this study assumes a generic nonlinear state space representation

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1},\, u_k) + \boldsymbol{v}_k \tag{2}$$

$$\boldsymbol{y}_k = g(\boldsymbol{x}_k) + \boldsymbol{n}_k, \tag{3}$$

where $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ denote the hidden state and measurement vectors and u_k is a scalar control input at discrete time index k. The system dynamics and measurements are described by the nonlinear functions $f(\cdot)$

and $g(\cdot)$, which are assumed to be time invariant. Process and measurement noise characteristics are modeled as additive, zero-mean Gaussian random variables $\boldsymbol{v}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ and $\boldsymbol{n}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$ with covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$.

System dynamics

The system dynamics introduced in Eq. (2) are represented as a 2-dimensional state vector $\boldsymbol{x}_k = \begin{bmatrix}\phi_k & \psi_k\end{bmatrix}^T$, where ϕ_k and ψ_k denote the angular direction of the sound source and the look direction of the head, respectively. In this study, the sound source position is assumed to be static, whereas head rotations are permitted within a restricted angular range $\psi_k \in [\psi_1, \psi_2]$. The system dynamics is modeled as

$$f(\boldsymbol{x}_{k-1}, u_k) = \begin{bmatrix} \phi_{k-1} \\ \mathrm{sat}(\psi_{k-1} + T\dot{\psi}_{\max} u_k, \psi_1, \psi_2) \end{bmatrix}, \quad (4)$$

where T denotes the step size between two consecutive discrete time steps, $\dot{\psi}_{\max}$ is the maximum angular velocity of the head rotation and $u_k \in [-1, 1]$ is the control input. To keep the look direction of the head within the specified angular range, a nonlinear saturation function

$$\mathrm{sat}(x, \psi_1, \psi_2) = \begin{cases} \psi_1, & \text{if } x < \psi_1 \\ x, & \text{if } \psi_1 \leq x \leq \psi_2 \\ \psi_2, & \text{if } x > \psi_2 \end{cases}$$

is used in Eq. (4). The initial state of the model is preassigned to $\boldsymbol{x}_0 = \begin{bmatrix}0 & \frac{\pi}{2}\end{bmatrix}^T$ in this study.

Measurement model

The measurement equation in Eq. (3) introduces a nonlinear mapping function $g(\boldsymbol{x}_k)$ from states $\boldsymbol{x}_k$ to observations $\boldsymbol{y}_k$. As described previously, ITDs and ILDs are used as primary binaural cues for the proposed localization model. It is assumed that the look direction of the head can directly be measured from the actuator position. Hence, the observations are modeled as a 3-dimensional vector $\boldsymbol{y}_k = \begin{bmatrix}\tau_k & \delta_k & \psi_k\end{bmatrix}^T$. To account for the circular nature of angular source positions and the look direction of the head, a regression model based on trigonometric functions

$$g(\boldsymbol{x}_k) = \begin{bmatrix} w_0^\tau + \sum_{n=1}^{N} w_n^\tau \sin\Big(n \cdot (\phi_k - \psi_k)\Big) \\ w_0^\delta + \sum_{n=1}^{N} w_n^\delta \sin\Big(n \cdot (\phi_k - \psi_k)\Big) \\ \psi_k \end{bmatrix} \quad (5)$$

is proposed for the prediction of ITDs and ILDs. The corresponding regression coefficients are denoted as w_n^τ and w_n^δ, respectively. The difference $\phi_k - \psi_k$ describes the relative source angle between the look direction of the head and the absolute source direction. The model in Eq. (5) resembles a finite Fourier series of order N with all cosine terms set to zero.

As the model is linear in the weights, they can be computed via supervised learning using conventional linear regression techniques [REF]. In this study, a training dataset is generated using anechoic head-related impulse responses (HRIRs) of the Knowles Electronics Manikin for Acoustic Research (KEMAR) dummy head [9], using white noise as stimulus signal. The dataset consists of binaural features extracted by the previously introduced binaural front-end at 360 different angular positions with an increment of 1° over the entire horizontal plane. The order of the Fourier series was determined as $N = 8$ using cross-validation, to closely match the HRIRs of the dummy head. Fig. 1 shows the predicted ITDs and ILDs using the trained model compared to the measured binaural cues of the KEMAR head.

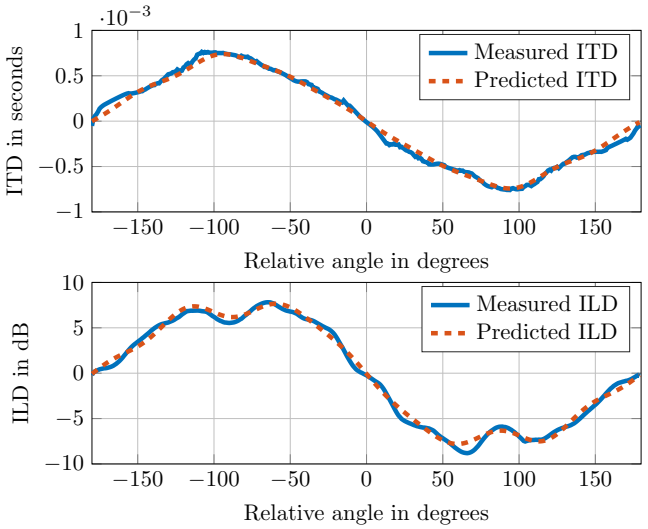

Figure 1: Comparison of measured and predicted ITD and ILD using the proposed linear regression model.

The measurement noise covariance matrix of the model introduced in Eq. (5) is assumed to be a block-diagonal matrix

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{\Sigma}_{\tau\delta} & \boldsymbol{0} \\ \boldsymbol{0}^T & \sigma_\psi^2 \end{bmatrix}, \quad (6)$$

where $\boldsymbol{\Sigma}_{\tau\delta}$ is the noise covariance matrix of the binaural features and σ_ψ^2 is the scalar measurement noise variance of the look direction. The former is estimated using the trained model (5) and the measured binaural cues from the training dataset. The measurement noise variance of the look direction is set to a fixed value of $\sigma_\psi^2 = 0.01$.

State estimation and feedback strategies

State estimation is performed with a generic unscented Kalman filter (UKF), using the publicly available EKF/UKF Toolbox [10]. The UKF computes an estimate of the state $\hat{\boldsymbol{x}}_k = \begin{bmatrix}\hat{\phi}_k & \hat{\psi}_k\end{bmatrix}^T$ with the corresponding estimation error covariance matrix $\hat{\boldsymbol{P}}_k$ at each time step. Throughout all experiments conducted in this study, the UKF is initialized with

$$\hat{\boldsymbol{x}}_0 = \begin{bmatrix}0 \\ 0\end{bmatrix} \quad \hat{\boldsymbol{P}}_0 = \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}.$$

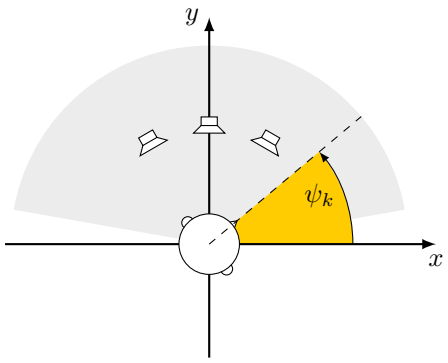

Figure 2: Arrangement of sound sources in the evaluation scenario using BRIRs of the room "Spirit". The scenario includes three sound sources, positioned in the frontal hemisphere of the dummy head. The gray area indicates the region of possible head rotations. For a detailed description of the source positions see Tab. 1.

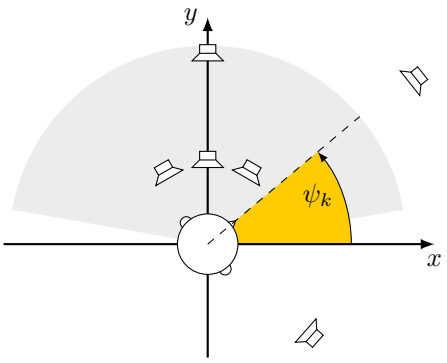

Figure 3: Arrangement of sound sources in the evaluation scenario using BRIRs of the room "Auditorium 3". The scenario includes six sound sources, of which five are positioned in the frontal hemisphere of the dummy head and one is positioned in the back.

The proposed systems allows the integration of feedback through continuous head rotations, based on the estimated state of the system. A positive control input of $u_k = 1$ triggers a clockwise head rotation with maximum angular velocity. Similarly, negative values induce a counter-clockwise head rotation. Based on previous work presented in [5], this study investigates the effect of two novel head rotation strategies on localization performance.

Static head position: No external control input is applied here. The control input is set to $u_k = 0\ \forall k$, hence the look direction of the dummy head will remain in its initial position. **Smooth posterior mean:** The smooth posterior mean (SPM) feedback strategy was introduced in [5] and is based on a closed-loop feedback control paradigm

$$u_k = \frac{|\hat{\phi}_k - \hat{\psi}_k|}{1 + |\hat{\phi}_k - \hat{\psi}_k|} \cdot \operatorname{sgn}\Big(\hat{\phi}_k - \hat{\psi}_k\Big). \tag{7}$$

This approach steers the look direction of the head on a smooth trajectory towards the estimated angular position of the sound source.

Proportional controller: The first feedback strategy proposed in this study is a saturated proportional controller

$$u_k = \operatorname{sat}\Big(\kappa_\mathrm{p}(\hat{\phi}_k - \hat{\psi}_k),\ -1,\ 1\Big) \tag{8}$$

which steers the look direction of the head towards the estimated source position. The proportional gain was determined empirically to $\kappa_\mathrm{p} = 1,6$ using the Ziegler-Nichols tuning method [11]. The saturation function is necessary to keep the control input within the valid range of $u_k \in [-1,\ 1]$, so that the maximum rotational velocity of the head can not be exceeded.

Extended proportional controller: The second feedback strategy is an extension of the previously defined proportional controller, which uses cycles of alternating feed-forward and feedback control schemes. Initially, the head is rotated clockwise with a fixed control input of $u_k = -u_\mathrm{FF}$ in an open-loop for a predefined amount of time steps K_FF. Empirical tests have shown that this helps the system to explore the state-space, which decreases the effect of possible front-back ambiguities. Subsequently, the controller switches to a cyclic alternation

$$u_k = \begin{cases} \operatorname{sat}\Big(\tilde{\kappa}_{\mathrm{p},k}(\hat{\phi}_k - \hat{\psi}_k),\ -1,\ 1\Big) & \text{if } (k \bmod K_\mathrm{FB}) < \frac{K_\mathrm{FB}}{2} \\ 0 & \text{otherwise} \end{cases}$$

where $\tilde{\kappa}_{\mathrm{p},k} = \frac{\mathrm{tr}(\hat{\boldsymbol{P}}_k)}{\mathrm{tr}(\hat{\boldsymbol{P}}_0)}$ is the adaptive proportional control gain and K_FB is the number of time steps in each cycle. Here, the adaptive gain factor explicitly takes into account the uncertainty of the estimated state by computing the trace of the estimation error covariance matrix. The normalization factor $\mathrm{tr}(\hat{\boldsymbol{P}}_0)$ is necessary to prevent $\tilde{\kappa}_{\mathrm{p},k}$ from getting too small.

Evaluation

The proposed system is evaluated in two different scenarios using simulated reverberant conditions. The freely available database introduced in [3] is used for this purpose. The database contains sets of BRIRs, recorded with a KEMAR dummy head in two different rooms, referred to as "Spirit" and "Auditorium 3". The estimated reverberation times of the rooms are $T_{60} \approx 0.5\,\mathrm{s}$ ("Spirit") and $T_{60} \approx 0.7\,\mathrm{s}$ ("Auditorium 3").

Experimental setup

Both evaluation scenarios incorporate fixed source positions at different azimuth angles with respect to the dummy head. An overview of the specific experimental setup is depicted in Figs. 2 and 3. The exact source positions within both rooms are listed in Tab. 1. Three different source positions were evaluated in the "Spirit" scenario. The evaluation of "Auditorium 3" incorporates six different source positions. The angular range of possible look directions is fixed at $\psi_k \in [10°,\ 170°]$ for all experiments.

The target sounds used in this study were taken from the "Detection and Classification of Acoustic Scenes and

Table 1: Evaluation results for both sets of BRIRs that were used in the experiments. The table shows localization errors in degrees for all source directions and feedback strategies that were considered during the evaluation. Bold numbers indicate the best result at each source direction.

	Source azimuth in degrees	Source distance in meters	Static head position	SPM controller	Proportional controller	Extended proportional controller
Auditorium 3	90.00	3.97	**3.38**	7.00	7.04	3.60
	38.49	5.5	42.40	8.60	**8.34**	8.75
	-41.40	2.67	127.71	31.84	30.70	**30.43**
	90.00	1.80	2.67	3.90	3.92	**2.50**
	120.00	1.80	8.85	4.26	4.28	**2.77**
	60.00	1.80	13.30	6.63	6.61	**5.45**
Spirit	120.00	2.00	16.83	17.92	18.84	**8.07**
	90.00	2.00	**4.91**	13.22	13.37	5.15
	60.00	2.00	27.32	20.89	20.87	**12.31**

Events" dataset [12], which contains sounds of 16 different acoustic events recorded in an office environment. Five sounds of each event class were chosen randomly during the evaluation. Silence periods were excluded using the corresponding alignments provided with the database. The individual sounds were replicated to match a total duration of 30 seconds, yielding a total amount of 40 minutes of audio material evaluated at each source position and for each feedback strategy in both rooms. Localization performance was measured for each sound file using the circular root mean square error (RMSE) [5]. As the state estimation using the UKF takes a certain amount of time to converge, a grace period is used during each simulation, excluding the first 5 seconds of the simulated signals from the performance assessment.

Results and discussion

The outcome of the evaluation procedure is summarized in Tab. 1. The results indicate that the investigated close-loop feedback approaches generally outperform the feed-forward case where no rotational head movements are conducted. This is especially evident for the source positioned at $-41.40°$ in the "Auditorium 3" scenario, where front-back ambiguities lead to a large localization error in the static case. Furthermore, the extended proportional controller proposed in this study shows superior performance compared to the conventional proportional controller and the SPM method introduced in [5]. The conducted experiments have revealed that an adaptation of the control gain dependent on the estimation uncertainty in combination with cyclic interruptions of the rotational movement lead to more accurate estimation results. In addition, the results show that the system is able to achieve considerable localization performance in reverberant environments.

Acknowledgements

This research has been supported by EU FET grant Two!Ears, ICT-618075.

References

[1] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, 1940.

[2] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1999.

[3] N. Ma, T. May, H. Wierstorf, and G. J. Brown, "A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2699–2703.

[4] N. Ma, G. J. Brown, and J. A. Gonzalez, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10*, 2015, pp. 3066–3070.

[5] C. Schymura, F. Winter, D. Kolossa, and S. Spors, "Binaural sound source localisation and tracking using a dynamic spherical head model," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10*, 2015, pp. 165–169.

[6] T. May, S. van de Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[7] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[8] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system: I. Model structure," *Journal of the Acoustical Society of America*, vol. 99, pp. 3615–3622, 1996.

[9] H. Wierstorf, M. Geier, and S. Spors, "A Free Database of Head Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances," in *Proc. of 130th Aud. Eng. Soc. Conv.*, London, UK, 2011.

[10] S. Särkkä, J. Hartikainen, and A. Solin, "EKF/UKF toolbox for Matlab V1. 3," Aug. 2011. [Online]. Available: http://becs.aalto.fi/en/research/bayes/ekfukf/

[11] J. G. Ziegler and N. B. Nichols, "Optimum Settings for Automatic Controllers," *Transactions of ASME*, vol. 64, pp. 759–768, 1942.

[12] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.