RUHR-UNIVERSITÄT BOCHUM

Faculty of Electrical Engineering and Information Technology
Cognitive Signal Processing Group

A Blackboard System for Binaural Sound Source Localization and Tracking

C. Schymura and D. Kolossa

Introduction

The human auditory system is capable of precisely locating static and moving sound sources, even in adverse acoustic environments. In this work, we propose a framework for mimicking this ability by computational means using a blackboard system. It is based on recursive Bayesian estimation via unscented Kalman filtering and incorporates hypothesis-driven feedback through different head rotation strategies. The results indicate that feedback significantly improves localization accuracy in comparison to conventional feed-forward approaches.

Binaural Front-End

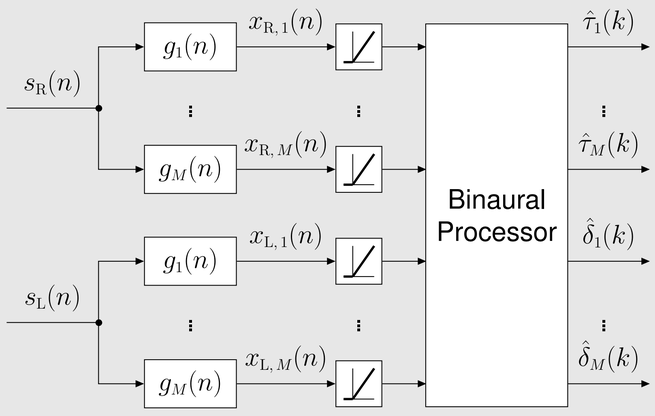

FIGURE 1: Overview of the binaural front-end used in this study.

n: Discrete time index
$s_{\{\mathrm{R,L}\}}(n)$: Audio signals at right and left ear
M: Number of filterbank channels
$g_i(n)$: i-th channel gammatone filter
$x_{\{\mathrm{R,L}\},i}(n)$: Filtered signal at i-th channel
k: Discrete frame index
$\hat{\tau}_i(k)$: Interaural time differences (ITDs)
$\hat{\delta}_i(k)$: Interaural level differences (ILDs)

At each frame, the binaural processor computes the interaural cross-correlation

$$c_i(k,\,\lambda) = \sum_{l=0}^{L-1} x_{\mathrm{L},i}(k\cdot L - l)\, x_{\mathrm{R},i}(k\cdot L - l - \lambda),$$

where L and λ are the frame length and time lag in samples, respectively. The ITD at each filterbank channel is estimated as

$$\hat{\tau}_i(k) = \frac{1}{f_{\mathrm{s}}}\left[\arg\max_{\lambda} c_i(k,\,\lambda)\right],$$

where f_{s} is the sampling frequency. The ILD is derived by comparing the energy of both ear signals

$$\hat{\delta}_i(k) = 10\,\log_{10}\left(\frac{\sum_{l=0}^{L-1} x_{\mathrm{R},i}(k\cdot L - l)^2}{\sum_{l=0}^{L-1} x_{\mathrm{L},i}(k\cdot L - l)^2}\right).$$

Blackboard System

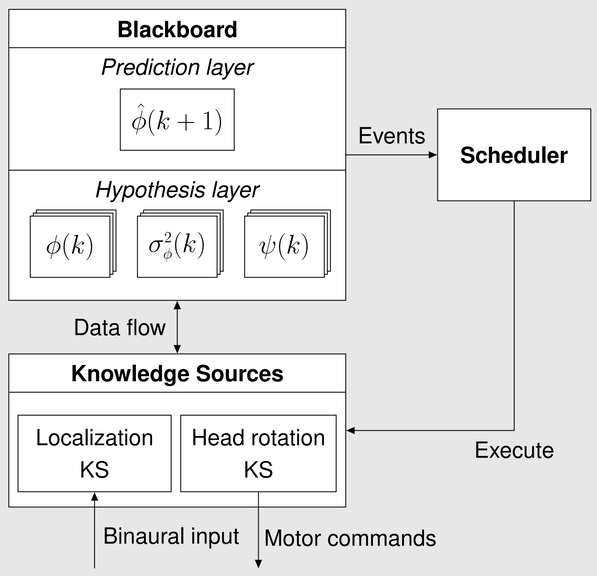

FIGURE 2: Architecture of the proposed blackboard system.

$\phi(k)$: Source azimuth hypothesis
$\sigma^2_\phi(k)$: Localization error variance
$\psi(k)$: Look direction hypothesis
$\hat{\phi}(k+1)$: Estimated source azimuth

The blackboard serves as a data repository, where hypotheses about the source location, the look direction of the head and the corresponding error variances are stored. Knowledge Sources (KSs) are independent functional blocks which can access and modify the data on the blackboard. The execution of individual KSs is controlled via a scheduling mechanism.

Knowledge Sources

- **Localization KS**: Performs localization and tracking of a sound source based on the nonlinear model dynamics

$$\begin{aligned}\phi(k+1) &= \phi(k) + \frac{L}{f_{\mathrm{s}}}\dot{\phi}(k) + v_\phi(k)\\ \dot{\phi}(k+1) &= \dot{\phi}(k) + v_{\dot{\phi}}(k)\\ \psi(k+1) &= \mathrm{sat}\left(\psi(k) + \frac{L}{f_{\mathrm{s}}}\dot{\psi}_{\max}u(k)\right) + v_\psi(k),\end{aligned}$$

presented in [1], where $\phi(k)$ is the source azimuth, $\dot{\phi}(k)$ is the angular velocity of the source, $\psi(k)$ is the head orientation and $\dot{\psi}_{\max}$ is the maximum angular velocity for rotating the head. The process noise variables $v_i(k)$ are assumed to be Gaussian. The control input $u(k)$ is restricted to $[-1,\,1]$. A saturation function $\mathrm{sat}(x) = \min(|x|,\,x_{\max})\cdot\mathrm{sgn}(x)$ models the constraints of maximum head rotation. The model dynamics incorporate the nonlinear observation equation

$$\boldsymbol{y}(k) = \boldsymbol{h}\Big(\phi(k),\,\psi(k)\Big) + \boldsymbol{w}(k).$$

The observations are composed of ITDs and ILDs estimated by the binaural front-end:

$$\boldsymbol{y}(k) = [\tau_1(k),\,\ldots,\,\tau_M(k),\,\delta_1(k),\,\ldots,\,\delta_M(k)]^T$$

and the Gaussian distributed noise variable $\boldsymbol{w}(k)$. The nonlinear function $\boldsymbol{h}(\cdot)$ provides a mapping from the relative location of a sound source $\phi_{\mathrm{r}}(k) = \phi(k) - \psi(k)$ to ITDs and ILDs using a spherical head model [2]. An unscented Kalman filter [3] is used to infer the angular position, velocity and head orientation. The Localization KS is executed once at each frame by the scheduling mechanism.

- **Head rotation KS**: Rotates the head on a smooth trajectory towards the estimated source position by generating the control input

$$u(k) = \left(1 - \frac{1}{1+|\phi_{\mathrm{r}}(k)|}\right)\cdot\mathrm{sgn}\Big(\phi_{\mathrm{r}}(k)\Big).$$

The Head rotation KS is executed if the prediction error variance $\sigma^2_\phi(k)$ exceeds a specified threshold value ϵ_σ.

Experimental Results

Monte-Carlo simulations with moving speech sources, starting at five different initial positions were performed. All sources were placed on a 3m radius from the head in anechoic conditions. The circular RMSE over all frames was used as an evaluation metric. Significant improvements can be achieved for all possible source positions.

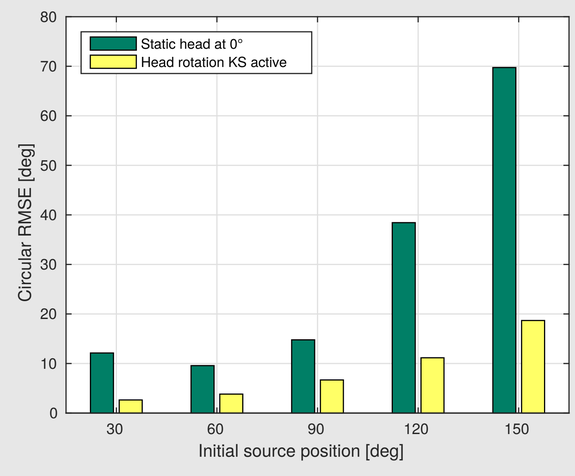

FIGURE 3: Evaluation results with and without head rotations.

References

[1] C. Schymura, F. Winter, D. Kolossa, and S. Spors, "Binaural Sound Source Localisation and Tracking using a Dynamic Spherical Head Model," in *INTERSPEECH 2015, Dresden, Germany, September 6-10*, 2015.

[2] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479, 1999.

[3] S. J. Julier and J. K. Uhlmann, "A New Extension of the Kalman Filter to Nonlinear Systems," *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*, vol. 3, 1997.