Inference in Nonlinear Dynamical Systems with Dynamic Stream Weights for Audiovisual Speaker Tracking

ICA 2019

Christopher Schymura and Dorothea Kolossa

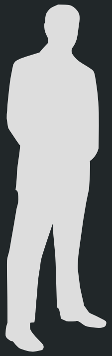
September 11th, 2019

RUHR UNIVERSITÄT BOCHUM

RUB

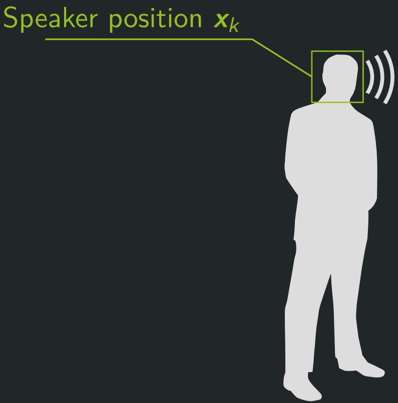
ICA 2019 AACHEN

Audiovisual speaker tracking

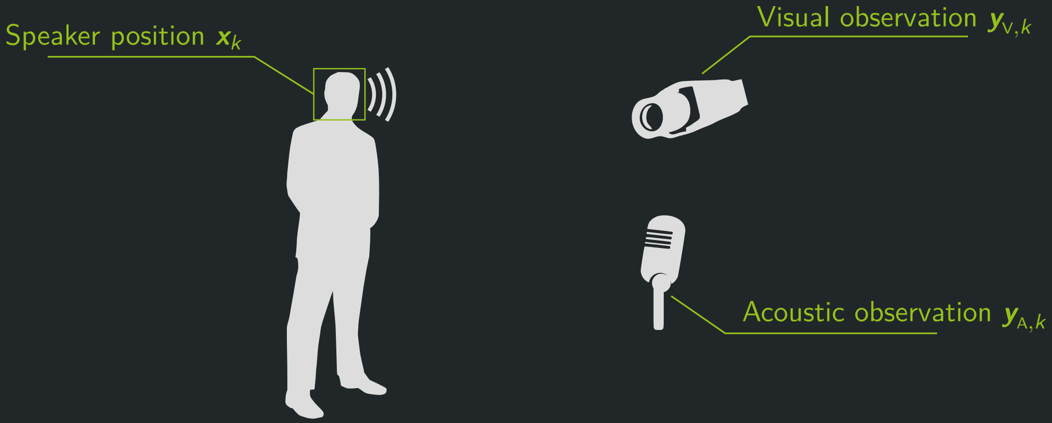


Audiovisual speaker tracking

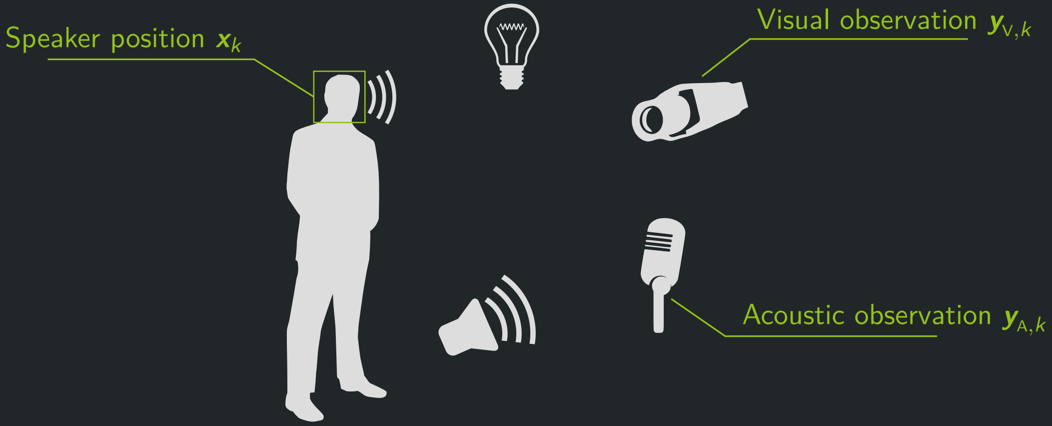
Audiovisual speaker tracking



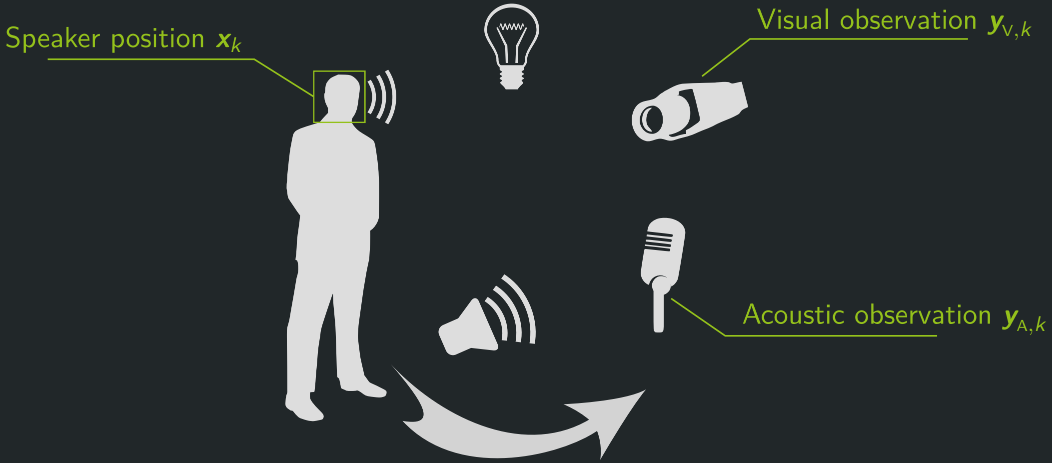
Audiovisual speaker tracking



Audiovisual speaker tracking



Audiovisual speaker tracking

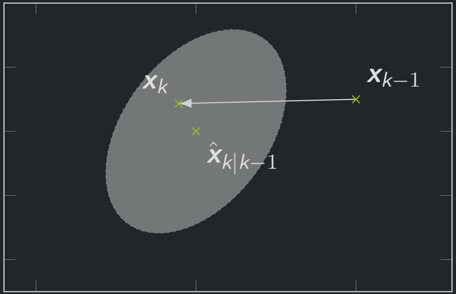


Audiovisual speaker tracking

Prediction step

System dynamics:

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_k, \quad \boldsymbol{v}_k = \mathcal{N}(\mathbf{0}, \boldsymbol{Q})$$

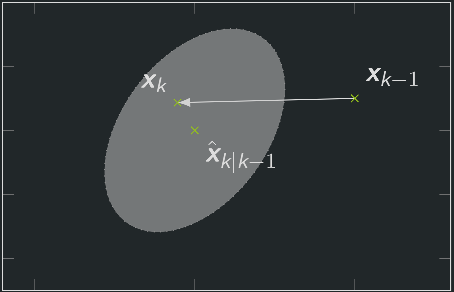


Audiovisual speaker tracking

Prediction step

System dynamics:

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_k, \quad \boldsymbol{v}_k = \mathcal{N}(\mathbf{0},\, \boldsymbol{Q})$$



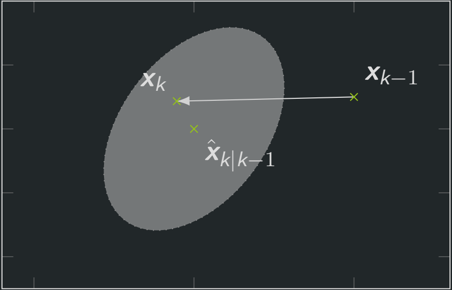
$$p(\boldsymbol{x}_k \,|\, \boldsymbol{Y}_{\mathrm{A},k-1},\, \boldsymbol{Y}_{\mathrm{V},k-1}) = \int \, p(\boldsymbol{x}_k \,|\, \boldsymbol{x}_{k-1}) \;\, p(\boldsymbol{x}_{k-1} \,|\, \boldsymbol{Y}_{\mathrm{A},k-1},\, \boldsymbol{Y}_{\mathrm{V},k-1}) \;\, d\boldsymbol{x}_{k-1}$$

Audiovisual speaker tracking

Prediction step

System dynamics:

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_k, \quad \boldsymbol{v}_k = \mathcal{N}(\mathbf{0},\ \boldsymbol{Q})$$



$$p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1}) = \int \underbrace{p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1})}_{\text{Dynamic model}} \underbrace{p(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1})}_{\text{Prior}} \, d\boldsymbol{x}_{k-1}$$

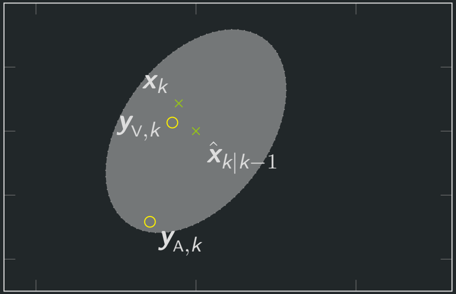
Audiovisual speaker tracking

Observation

Observation model:

$$\boldsymbol{y}_k = \begin{bmatrix}\boldsymbol{y}_{\mathrm{A},k} & \boldsymbol{y}_{\mathrm{V},k}\end{bmatrix}^{\mathsf{T}} = h\left(\boldsymbol{x}_k\right) + \boldsymbol{w}_k$$

$$\boldsymbol{w}_k = \mathcal{N}(0,\ \boldsymbol{R}), \quad \boldsymbol{R} = \begin{bmatrix}\boldsymbol{R}_{\mathrm{AA}} & \boldsymbol{R}_{\mathrm{AV}} \\ \boldsymbol{R}_{\mathrm{VA}} & \boldsymbol{R}_{\mathrm{VV}}\end{bmatrix}$$



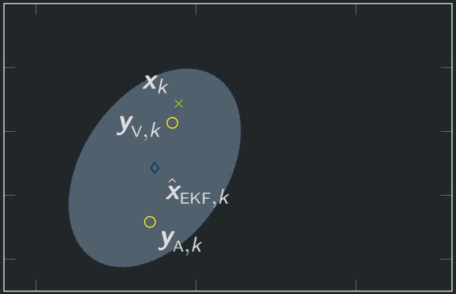
Audiovisual speaker tracking

Update step (standard Kalman filter)

Observation model:

$$\mathbf{y}_k = \begin{bmatrix} \mathbf{y}_{\text{A},k} & \mathbf{y}_{\text{V},k} \end{bmatrix}^{\text{T}} = h\left(\mathbf{x}_k\right) + \mathbf{w}_k$$

$$\mathbf{w}_k = \mathcal{N}(0,\ \mathbf{R}), \quad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{\text{AA}} & \mathbf{R}_{\text{AV}} \\ \mathbf{R}_{\text{VA}} & \mathbf{R}_{\text{VV}} \end{bmatrix}$$



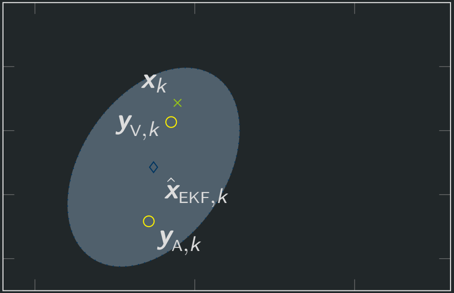
Audiovisual speaker tracking

Update step (standard Kalman filter)

Observation model:

$$\boldsymbol{y}_k = \begin{bmatrix} \boldsymbol{y}_{\mathrm{A},k} & \boldsymbol{y}_{\mathrm{V},k} \end{bmatrix}^{\mathsf{T}} = h\left(\boldsymbol{x}_k\right) + \boldsymbol{w}_k$$

$$\boldsymbol{w}_k = \mathcal{N}(0,\ \boldsymbol{R}), \quad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{\mathrm{AA}} & \boldsymbol{R}_{\mathrm{AV}} \\ \boldsymbol{R}_{\mathrm{VA}} & \boldsymbol{R}_{\mathrm{VV}} \end{bmatrix}$$



$$p(\boldsymbol{x}_k \,|\, \boldsymbol{Y}_{\mathrm{A},k},\ \boldsymbol{Y}_{\mathrm{V},k}) \propto p(\boldsymbol{x}_k \,|\, \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1})\, p(\boldsymbol{y}_{\mathrm{A},k},\ \boldsymbol{y}_{\mathrm{V},k} \,|\, \boldsymbol{x}_k)$$

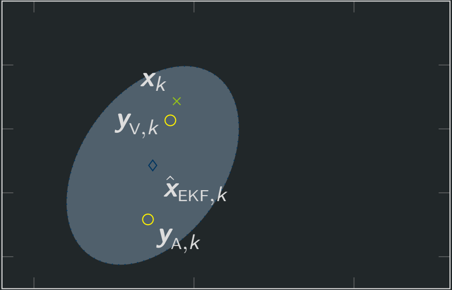
Audiovisual speaker tracking

Update step (standard Kalman filter)

Observation model:

$$\boldsymbol{y}_k = \begin{bmatrix} \boldsymbol{y}_{\mathrm{A},k} & \boldsymbol{y}_{\mathrm{V},k} \end{bmatrix}^{\mathsf{T}} = h\left(\boldsymbol{x}_k\right) + \boldsymbol{w}_k$$

$$\boldsymbol{w}_k = \mathcal{N}(0,\ \boldsymbol{R}), \quad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{\mathrm{AA}} & \boldsymbol{R}_{\mathrm{AV}} \\ \boldsymbol{R}_{\mathrm{VA}} & \boldsymbol{R}_{\mathrm{VV}} \end{bmatrix}$$



$$p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{\mathrm{A},k},\ \boldsymbol{Y}_{\mathrm{V},k}) \propto p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1}) \underbrace{p(\boldsymbol{y}_{\mathrm{A},k},\ \boldsymbol{y}_{\mathrm{V},k} \mid \boldsymbol{x}_k)}_{\text{Sensor model}}$$

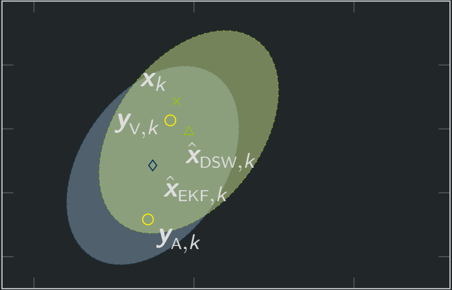
Audiovisual speaker tracking

Update step (Kalman filter with dynamic stream weights[1])

Observation model:

$$\boldsymbol{y}_{\mathrm{A},k} = h_{\mathrm{A}}(\boldsymbol{x}_k) + \boldsymbol{w}_{\mathrm{A},k}, \quad \boldsymbol{w}_{\mathrm{A},k} = \mathcal{N}(0,\ \boldsymbol{R}_{\mathrm{AA}})$$

$$\boldsymbol{y}_{\mathrm{V},k} = h_{\mathrm{V}}(\boldsymbol{x}_k) + \boldsymbol{w}_{\mathrm{V},k}, \quad \boldsymbol{w}_{\mathrm{V},k} = \mathcal{N}(0,\ \boldsymbol{R}_{\mathrm{VV}})$$



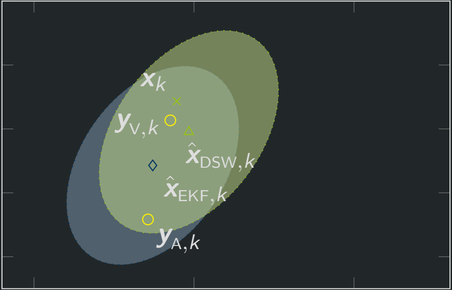
[1] C. Schymura et al.: *Extending linear dynamical systems with dynamic stream weights for audiovisual speaker localization*, IWAENC, 2018

Audiovisual speaker tracking

Update step (Kalman filter with dynamic stream weights[1])

Observation model:

$$\boldsymbol{y}_{\mathrm{A},k} = h_{\mathrm{A}}(\boldsymbol{x}_k) + \boldsymbol{w}_{\mathrm{A},k}, \quad \boldsymbol{w}_{\mathrm{A},k} = \mathcal{N}(0,\ \boldsymbol{R}_{\mathrm{AA}})$$
$$\boldsymbol{y}_{\mathrm{V},k} = h_{\mathrm{V}}(\boldsymbol{x}_k) + \boldsymbol{w}_{\mathrm{V},k}, \quad \boldsymbol{w}_{\mathrm{V},k} = \mathcal{N}(0,\ \boldsymbol{R}_{\mathrm{VV}})$$



$$p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{\mathrm{A},k},\ \boldsymbol{Y}_{\mathrm{V},k}) \propto p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1}) \underbrace{p(\boldsymbol{y}_{\mathrm{A},k} \mid \boldsymbol{x}_k)^{\lambda_k}}_{\text{Acoustic model}} \underbrace{p(\boldsymbol{y}_{\mathrm{V},k} \mid \boldsymbol{x}_k)^{1-\lambda_k}}_{\text{Visual model}}$$

[1] C. Schymura et al.: *Extending linear dynamical systems with dynamic stream weights for audiovisual speaker localization*, IWAENC, 2018

Inference

Extended Kalman filter approach: first-order Taylor series expansion

$$f(\boldsymbol{x}_{k-1}) \approx f(\hat{\boldsymbol{x}}_{k-1}) + \boldsymbol{F}(\hat{\boldsymbol{x}}_{k-1})(\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-1})$$

Inference

Extended Kalman filter approach: first-order Taylor series expansion

$$f(\boldsymbol{x}_{k-1}) \approx f(\hat{\boldsymbol{x}}_{k-1}) + \boldsymbol{F}(\hat{\boldsymbol{x}}_{k-1})(\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-1})$$

$$\Rightarrow \quad p(\boldsymbol{x}_k \,|\, \boldsymbol{x}_{k-1}) = \mathcal{N}\left(\boldsymbol{x}_k \,|\, f(\hat{\boldsymbol{x}}_{k-1}) + \boldsymbol{F}(\hat{\boldsymbol{x}}_{k-1})(\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-1}),\ \boldsymbol{Q}\right)$$

$$\Rightarrow \quad p(\boldsymbol{x}_k \,|\, \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1}) = \mathcal{N}\left(\boldsymbol{x}_k \,|\, \hat{\boldsymbol{x}}_{k-1},\ \hat{\Sigma}_{k-1}\right)$$

Inference

Extended Kalman filter approach: first-order Taylor series expansion

$$f(\boldsymbol{x}_{k-1}) \approx f(\hat{\boldsymbol{x}}_{k-1}) + \boldsymbol{F}(\hat{\boldsymbol{x}}_{k-1})(\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-1})$$

$$\Rightarrow \qquad p(\boldsymbol{x}_k \,|\, \boldsymbol{x}_{k-1}) = \mathcal{N}\Big(\boldsymbol{x}_k \,\Big|\, f(\hat{\boldsymbol{x}}_{k-1}) + \boldsymbol{F}(\hat{\boldsymbol{x}}_{k-1})(\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-1}),\ \boldsymbol{Q}\Big)$$

$$\Rightarrow \qquad p(\boldsymbol{x}_k \,|\, \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1}) = \mathcal{N}\Big(\boldsymbol{x}_k \,\Big|\, \hat{\boldsymbol{x}}_{k-1},\ \hat{\boldsymbol{\Sigma}}_{k-1}\Big)$$

Prediction step (identical to standard EKF)

$$\hat{\boldsymbol{x}}_{k|k-1} = f(\hat{\boldsymbol{x}}_{k-1})$$

$$\hat{\boldsymbol{\Sigma}}_{k|k-1} = \boldsymbol{F}_{k-1}\hat{\boldsymbol{\Sigma}}_{k-1}\boldsymbol{F}_{k-1}^{\mathsf{T}} + \boldsymbol{Q}, \qquad \boldsymbol{F}_{k-1} \equiv \boldsymbol{F}(\hat{\boldsymbol{x}}_{k-1}) = \frac{\partial f(\boldsymbol{x}_{k-1})}{\partial \boldsymbol{x}_{k-1}}\Big|_{\boldsymbol{x}_{k-1}=\hat{\boldsymbol{x}}_{k-1}}$$

Inference

Extended Kalman filter approach: first-order Taylor series expansion

$$h_{\{\mathrm{A,V}\}}(\boldsymbol{x}_k) \approx h_{\{\mathrm{A,V}\}}(\hat{\boldsymbol{x}}_k) + \boldsymbol{H}_{\{\mathrm{A,V}\},k}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k), \quad \boldsymbol{H}_{\{\mathrm{A,V}\},k} \equiv \left.\frac{\partial h_{\{\mathrm{A,V}\}}(\boldsymbol{x}_k)}{\partial \boldsymbol{x}_k}\right|_{\boldsymbol{x}_k = \hat{\boldsymbol{x}}_k}$$

Inference

Extended Kalman filter approach: first-order Taylor series expansion

$$h_{\{A,V\}}(\boldsymbol{x}_k) \approx h_{\{A,V\}}(\hat{\boldsymbol{x}}_k) + \boldsymbol{H}_{\{A,V\},k}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k), \quad \boldsymbol{H}_{\{A,V\},k} \equiv \left.\frac{\partial h_{\{A,V\}}(\boldsymbol{x}_k)}{\partial \boldsymbol{x}_k}\right|_{\boldsymbol{x}_k=\hat{\boldsymbol{x}}_k}$$

$$\Rightarrow \qquad p(\boldsymbol{y}_{\{A,V\},k} \,|\, \boldsymbol{x}_k) = \mathcal{N}\Big(\boldsymbol{y}_{\{A,V\},k}, \,|\, h_{\{A,V\}}(\hat{\boldsymbol{x}}_k) + \boldsymbol{H}_{\{A,V\},k})(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k), \, \boldsymbol{R}_{\{A,V\}}\Big)$$

Inference

Extended Kalman filter approach: first-order Taylor series expansion

$$h_{\{A,V\}}(\boldsymbol{x}_k) \approx h_{\{A,V\}}(\hat{\boldsymbol{x}}_k) + \boldsymbol{H}_{\{A,V\},k}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k), \quad \boldsymbol{H}_{\{A,V\},k} \equiv \left.\frac{\partial h_{\{A,V\}}(\boldsymbol{x}_k)}{\partial \boldsymbol{x}_k}\right|_{\boldsymbol{x}_k=\hat{\boldsymbol{x}}_k}$$

$$\Rightarrow \qquad p(\boldsymbol{y}_{\{A,V\},k} \,|\, \boldsymbol{x}_k) = \mathcal{N}\left(\boldsymbol{y}_{\{A,V\},k}; \,|\, h_{\{A,V\}}(\hat{\boldsymbol{x}}_k) + \boldsymbol{H}_{\{A,V\},k})(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k), \, \boldsymbol{R}_{\{A,V\}}\right)$$

Update step

$$\begin{bmatrix} \boldsymbol{K}_{A,k}^T \\ \boldsymbol{K}_{V,k}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_A + \lambda_k \boldsymbol{H}_{A,k}\hat{\boldsymbol{\Sigma}}_{k|k-1}\boldsymbol{H}_{A,k}^T & (1-\lambda_k)\boldsymbol{H}_{A,k}\hat{\boldsymbol{\Sigma}}_{k|k-1}\boldsymbol{H}_{V,k}^T \\ \lambda_k \boldsymbol{H}_{V,k}\hat{\boldsymbol{\Sigma}}_{k|k-1}\boldsymbol{H}_{A,k}^T & \boldsymbol{R}_V + (1-\lambda_k)\boldsymbol{H}_{V,k}\hat{\boldsymbol{\Sigma}}_{k|k-1}\boldsymbol{H}_{V,k}^T \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{H}_{A,k} \\ \boldsymbol{H}_{V,k} \end{bmatrix} \hat{\boldsymbol{\Sigma}}_{k|k-1} \tag{1}$$

$$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_{k|k-1} + \lambda_k \boldsymbol{K}_{A,k}\left(\boldsymbol{y}_{A,k} - h_A(\hat{\boldsymbol{x}}_k)\right) + (1-\lambda_k)\boldsymbol{K}_{V,k}\left(\boldsymbol{y}_{V,k} - h_V(\hat{\boldsymbol{x}}_k)\right)$$

$$\hat{\boldsymbol{\Sigma}}_{k|k-1} = \left(\boldsymbol{I} - \lambda_k \boldsymbol{K}_{A,k}\boldsymbol{H}_{A,k} - (1-\lambda_k)\boldsymbol{K}_{V,k}\boldsymbol{H}_{V,k}\right)\hat{\boldsymbol{\Sigma}}_{k|k-1}$$

Inference

The system of linear matrix equations in Eq. (1) can be expressed as

$$\begin{bmatrix} \boldsymbol{K}_{\mathrm{A},k}^{\mathsf{T}} & \boldsymbol{K}_{\mathrm{V},k}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{R} + \boldsymbol{U}_k \boldsymbol{W}_k \boldsymbol{U}_k^{\mathsf{T}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{H}_{\mathrm{A},k} & \boldsymbol{H}_{\mathrm{V},k} \end{bmatrix}^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}_{k|k-1}$$

$$\boldsymbol{R} = \mathrm{blkdiag}(\boldsymbol{R}_{\mathrm{A}},\ \boldsymbol{R}_{\mathrm{V}}),\ \ \boldsymbol{U}_k = \mathrm{blkdiag}(\boldsymbol{H}_{\mathrm{A},k},\ \boldsymbol{H}_{\mathrm{V},k}),\ \ \boldsymbol{W}_k = \begin{bmatrix} \lambda_k & 1-\lambda_k \\ \lambda_k & 1-\lambda_k \end{bmatrix} \otimes \hat{\boldsymbol{\Sigma}}_{k|k-1}$$

Inference

The system of linear matrix equations in Eq. (1) can be expressed as

$$\begin{bmatrix} \boldsymbol{K}_{\mathrm{A},k}^{\mathsf{T}} & \boldsymbol{K}_{\mathrm{V},k}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{R} + \boldsymbol{U}_k \boldsymbol{W}_k \boldsymbol{U}_k^{\mathsf{T}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{H}_{\mathrm{A},k} & \boldsymbol{H}_{\mathrm{V},k} \end{bmatrix}^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}_{k|k-1}$$

$$\boldsymbol{R} = \mathrm{blkdiag}(\boldsymbol{R}_{\mathrm{A}}, \ \boldsymbol{R}_{\mathrm{V}}), \ \ \boldsymbol{U}_k = \mathrm{blkdiag}(\boldsymbol{H}_{\mathrm{A},k}, \ \boldsymbol{H}_{\mathrm{V},k}), \ \ \boldsymbol{W}_k = \begin{bmatrix} \lambda_k & 1-\lambda_k \\ \lambda_k & 1-\lambda_k \end{bmatrix} \otimes \hat{\boldsymbol{\Sigma}}_{k|k-1}$$

Modified Kalman gain computation using the binomial inverse theorem[2]

$$\begin{bmatrix} \boldsymbol{K}_{\mathrm{A},k}^{\mathsf{T}} & \boldsymbol{K}_{\mathrm{V},k}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1} \boldsymbol{U}_k \boldsymbol{\Gamma}_k \boldsymbol{U}_k^{\mathsf{T}} \boldsymbol{R}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_{\mathrm{A},k} & \boldsymbol{H}_{\mathrm{V},k} \end{bmatrix}^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}_{k|k-1}, \quad \boldsymbol{\Gamma}_k = \boldsymbol{W}_k \left(\boldsymbol{I} + \boldsymbol{U}_k^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{U}_k \boldsymbol{W}_k \right)^{-1}$$

[2] D. Harville: *Extension of the Gauss-Markov theorem to include the estimation of random effects*, Ann. Statist. vol.4, no. 2, 1976

Inference

The system of linear matrix equations in Eq. (1) can be expressed as

$$\begin{bmatrix} \boldsymbol{K}_{\mathrm{A},k}^{\mathsf{T}} & \boldsymbol{K}_{\mathrm{V},k}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{R} + \boldsymbol{U}_k \boldsymbol{W}_k \boldsymbol{U}_k^{\mathsf{T}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{H}_{\mathrm{A},k} & \boldsymbol{H}_{\mathrm{V},k} \end{bmatrix}^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}_{k|k-1}$$

$$\boldsymbol{R} = \mathrm{blkdiag}(\boldsymbol{R}_{\mathrm{A}},\ \boldsymbol{R}_{\mathrm{V}}),\ \ \boldsymbol{U}_k = \mathrm{blkdiag}(\boldsymbol{H}_{\mathrm{A},k},\ \boldsymbol{H}_{\mathrm{V},k}),\ \ \boldsymbol{W}_k = \begin{bmatrix} \lambda_k & 1-\lambda_k \\ \lambda_k & 1-\lambda_k \end{bmatrix} \otimes \hat{\boldsymbol{\Sigma}}_{k|k-1}$$

Modified Kalman gain computation using the binomial inverse theorem[2]

$$\begin{bmatrix} \boldsymbol{K}_{\mathrm{A},k}^{\mathsf{T}} & \boldsymbol{K}_{\mathrm{V},k}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1} \boldsymbol{U}_k \boldsymbol{\Gamma}_k \boldsymbol{U}_k^{\mathsf{T}} \boldsymbol{R}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_{\mathrm{A},k} & \boldsymbol{H}_{\mathrm{V},k} \end{bmatrix}^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}_{k|k-1}, \quad \boldsymbol{\Gamma}_k = \boldsymbol{W}_k \left(\boldsymbol{I} + \boldsymbol{U}_k^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{U}_k \boldsymbol{W}_k \right)^{-1}$$

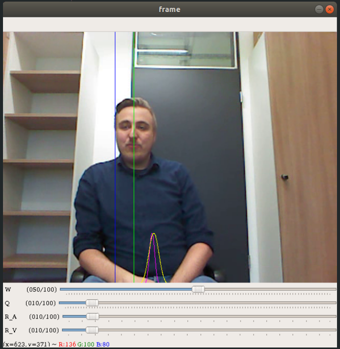
Complexity w.r.t. matrix inversions: $\mathcal{O}\left(8D_x^3\right)$ vs. $\mathcal{O}\left((D_{y_{\mathrm{A}}} + D_{y_{\mathrm{V}}})^3\right)$

[2] D. Harville: *Extension of the Gauss-Markov theorem to include the estimation of random effects*, Ann. Statist. vol.4, no. 2, 1976

Evaluation

Experimental setup

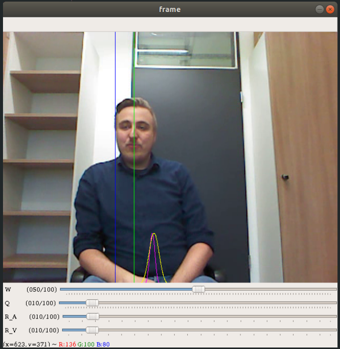
- KAVTraC audiovisual dataset, recorded in an office room at RUB using the Kinect sensor (7 speakers, $T_{60} \approx 350\,\text{ms}$, 35 min. duration).



Evaluation

Experimental setup

- KAVTraC audiovisual dataset, recorded in an office room at RUB using the Kinect sensor (7 speakers, $T_{60} \approx 350\,\mathrm{ms}$, 35 min. duration).
- Constant velocity linear dynamics model and nonlinear rotating vector observation models.
- DSW-EKF uses Dirichlet-prior oracle DSWs[3].

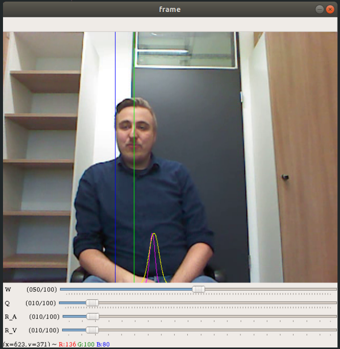


[3] C. Schymura et al.: *Audiovisual speaker tracking using nonlinear dynamical systems with dynamic stream weights*, arXiv, 2019

Evaluation

Experimental setup

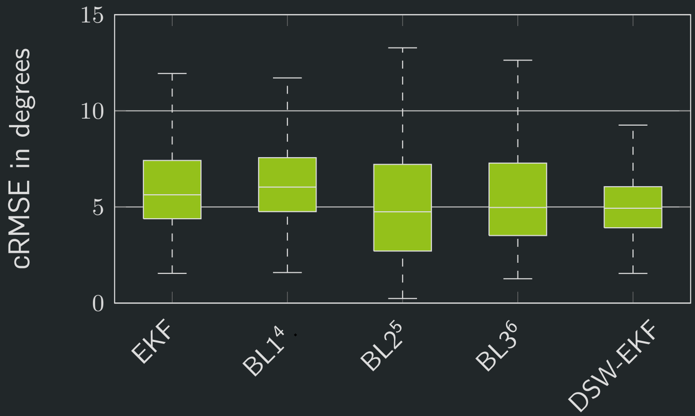
- KAVTraC audiovisual dataset, recorded in an office room at RUB using the Kinect sensor (7 speakers, $T_{60} \approx 350\,\text{ms}$, 35 min. duration).
- Constant velocity linear dynamics model and nonlinear rotating vector observation models.
- DSW-EKF uses Dirichlet-prior oracle DSWs[3].
- Four baseline systems: standard EKF, one KF-based and two particle filter-based systems.
- Leave-one-out cross-validation paradigm.



[3] C. Schymura et al.: *Audiovisual speaker tracking using nonlinear dynamical systems with dynamic stream weights*, arXiv, 2019

Evaluation

Results



[4] T. Gehrig et al.: *Kalman filters for audio-video source localization*, WASPAA, 2005

[5] S. Gerlach et al.: *2D audio-visual localization in home environments using a particle filter*, ITG Symp., 2012

[6] X. Qian et al.: *3D audio-visual speaker tracking with an adaptive particle filter*, ICASSP, 2017

Conclusions and outlook

- DSW-based audiovisual speaker tracking frameworks can be extended to cope with nonlinear systems.

Conclusions and outlook

- DSW-based audiovisual speaker tracking frameworks can be extended to cope with nonlinear systems.
- Complexity of update step adaptable to application.

Conclusions and outlook

- DSW-based audiovisual speaker tracking frameworks can be extended to cope with nonlinear systems.
- Complexity of update step adaptable to application.
- Ideas for future work:

Conclusions and outlook

- DSW-based audiovisual speaker tracking frameworks can be extended to cope with nonlinear systems.
- Complexity of update step adaptable to application.
- Ideas for future work:
  - Nonlinear dimensionality reduction for audiovisual observations via encoder networks.

Conclusions and outlook

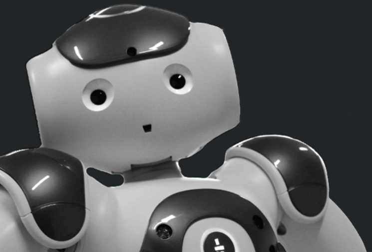
- DSW-based audiovisual speaker tracking frameworks can be extended to cope with nonlinear systems.
- Complexity of update step adaptable to application.
- Ideas for future work:
  - Nonlinear dimensionality reduction for audiovisual observations via encoder networks.
  - Joint optimization of model and DSW estimation in a deep learning framework.

Conclusions and outlook

- DSW-based audiovisual speaker tracking frameworks can be extended to cope with nonlinear systems.
- Complexity of update step adaptable to application.
- Ideas for future work:
  - Nonlinear dimensionality reduction for audiovisual observations via encoder networks.
  - Joint optimization of model and DSW estimation in a deep learning framework.
  - Extension to multi-speaker scenarios.

Conclusions and outlook

- DSW-based audiovisual speaker tracking frameworks can be extended to cope with nonlinear systems.
- Complexity of update step adaptable to application.
- Ideas for future work:
  - Nonlinear dimensionality reduction for audiovisual observations via encoder networks.
  - Joint optimization of model and DSW estimation in a deep learning framework.
  - Extension to multi-speaker scenarios.



Thank you for your attention!