Cognitive models for acoustic and audiovisual sound source localization

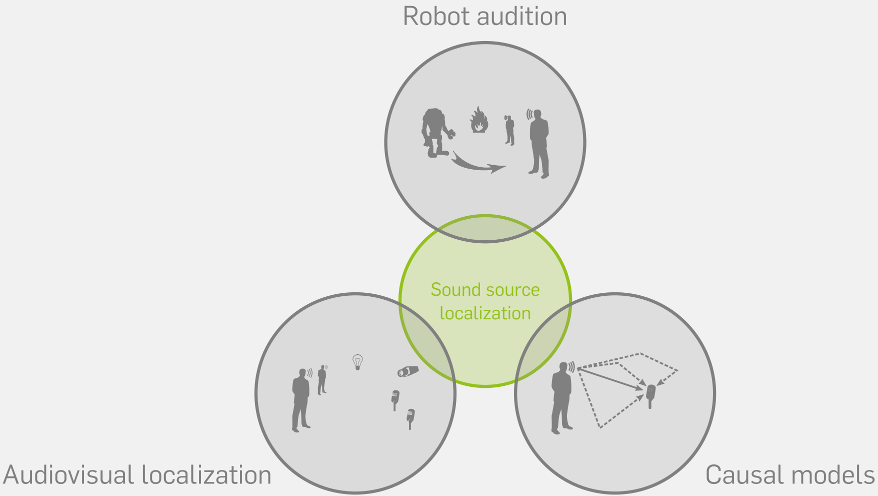
PhD thesis defense

Christopher Schymura

12th November 2019

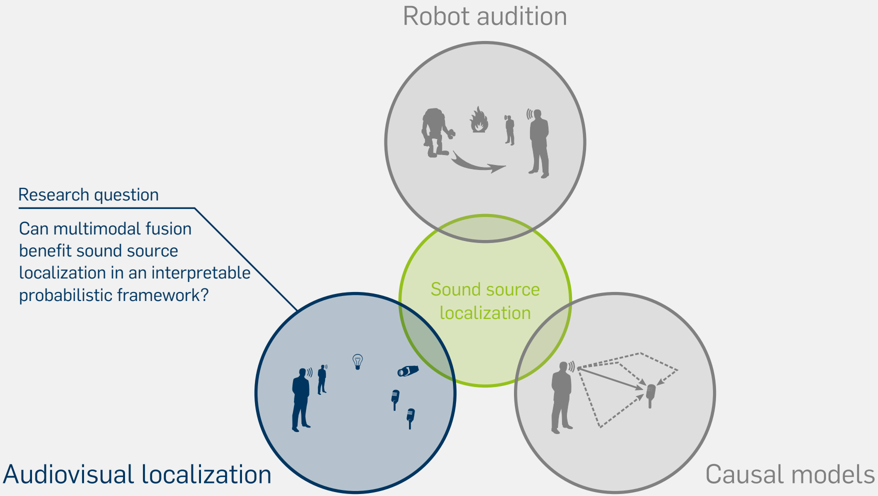
RUHR UNIVERSITÄT BOCHUM

Outline



Robot audition
Sound source localization
Audiovisual localization
Causal models

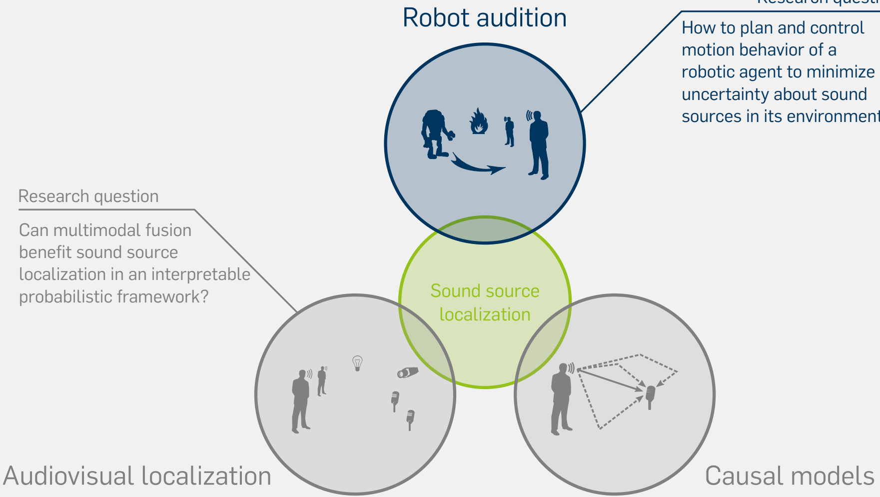
Outline


Robot audition
Research question
Can multimodal fusion benefit sound source localization in an interpretable probabilistic framework?
Sound source localization
Audiovisual localization
Causal models

Outline

Research question

How to plan and control motion behavior of a robotic agent to minimize uncertainty about sound sources in its environment?



Research question

Can multimodal fusion benefit sound source localization in an interpretable probabilistic framework?

Outline

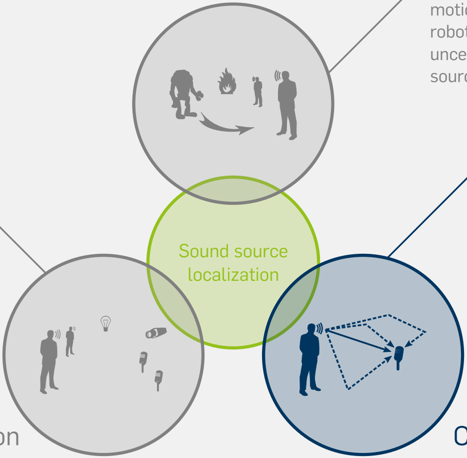
Robot audition

Research question

How to plan and control motion behavior of a robotic agent to minimize uncertainty about sound sources in its environment?

Research question

Can multimodal fusion benefit sound source localization in an interpretable probabilistic framework?



Research question

Can causal reasoning help to distinguish direct sound from reflections?

Audiovisual localization

Causal models

Outline

Robot audition

Research question

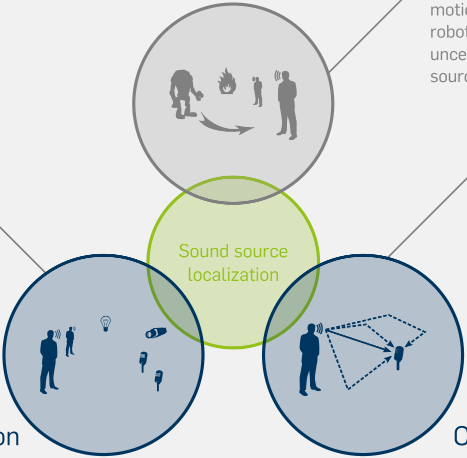
How to plan and control motion behavior of a robotic agent to minimize uncertainty about sound sources in its environment?

Research question

Can multimodal fusion benefit sound source localization in an interpretable probabilistic framework?

Research question

Can causal reasoning help to distinguish direct sound from reflections?



Audiovisual localization [Part I]

Causal models [Part II]

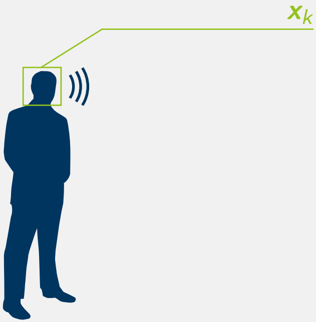
Part I

Audiovisual localization

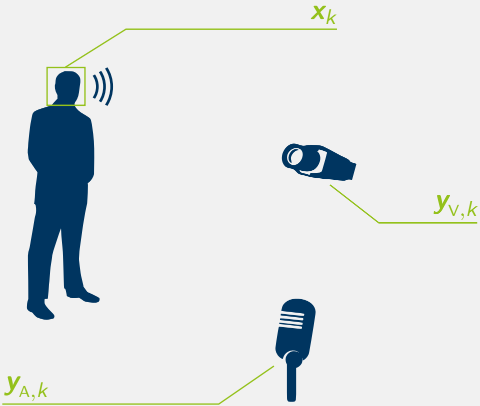
Audiovisual localization: Problem statement

Audiovisual localization: Problem statement

Audiovisual localization: Problem statement



Audiovisual localization: Problem statement

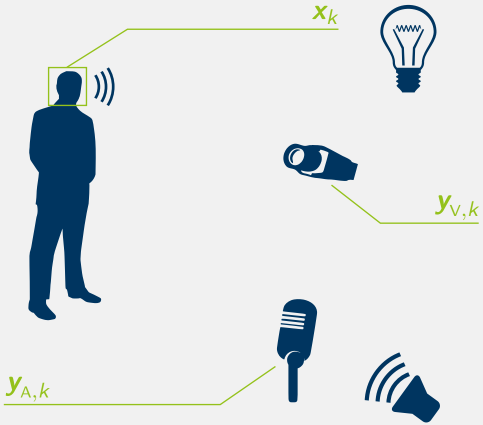


Observation functions:

$$\boldsymbol{y}_{\mathrm{A},k} = h_{\mathrm{A}}(\boldsymbol{x}_k)$$

$$\boldsymbol{y}_{\mathrm{V},k} = h_{\mathrm{V}}(\boldsymbol{x}_k)$$

Audiovisual localization: Problem statement

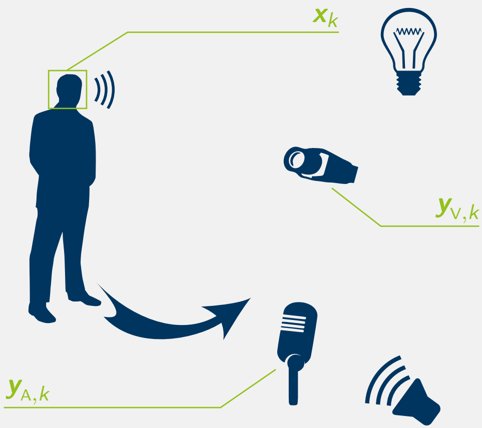


Observation functions:

$$\boldsymbol{y}_{\mathrm{A},k} = h_{\mathrm{A}}(\boldsymbol{x}_k) + \boldsymbol{w}_{\mathrm{A},k}$$

$$\boldsymbol{y}_{\mathrm{V},k} = h_{\mathrm{V}}(\boldsymbol{x}_k) + \boldsymbol{w}_{\mathrm{V},k}$$

Audiovisual localization: Problem statement



State transition function:

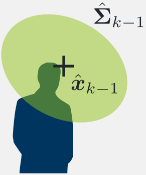
$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_k$$

Observation functions:

$$\boldsymbol{y}_{A,k} = h_A(\boldsymbol{x}_k) + \boldsymbol{w}_{A,k}$$

$$\boldsymbol{y}_{V,k} = h_V(\boldsymbol{x}_k) + \boldsymbol{w}_{V,k}$$

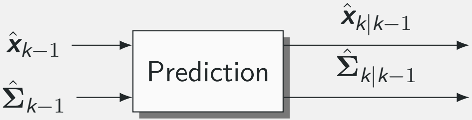
Audiovisual localization: Recursive state estimation



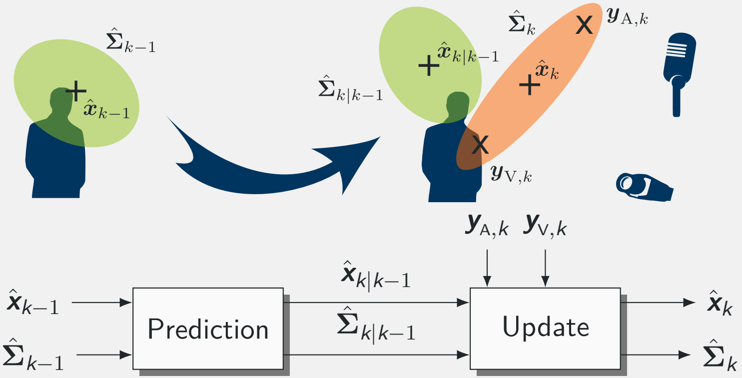
$\hat{\boldsymbol{x}}_{k-1}$

$\hat{\boldsymbol{\Sigma}}_{k-1}$

Audiovisual localization: Recursive state estimation

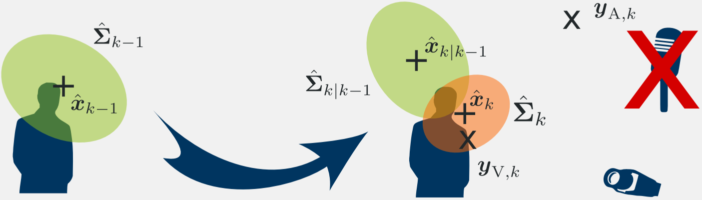


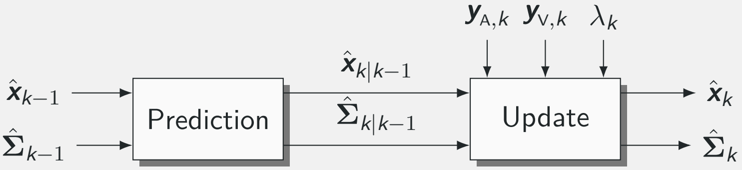
Audiovisual localization: Recursive state estimation



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

Audiovisual localization: Recursive state estimation





$$\underbrace{p(\boldsymbol{x}_k | \boldsymbol{Y}_{\mathrm{A},1:k}, \boldsymbol{Y}_{\mathrm{V},1:k})}_{\text{Posterior}} \propto \underbrace{p(\boldsymbol{x}_k | \boldsymbol{Y}_{\mathrm{A},1:k-1}, \boldsymbol{Y}_{\mathrm{V},1:k-1})}_{\text{Prior}} \underbrace{p(\boldsymbol{y}_{\mathrm{A},k} | \boldsymbol{x}_k)^{\lambda_k} \, p(\boldsymbol{y}_{\mathrm{V},k} | \boldsymbol{x}_k)^{1-\lambda_k}}_{\text{Sensor model w. stream weights}^2}$$

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

Audiovisual localization: Oracle dynamic stream weights

Assumption: $\boldsymbol{x}_k$, $\boldsymbol{y}_{\mathrm{A},k}$, $\boldsymbol{y}_{\mathrm{V},k}$, $k = 1, \ldots, K$ fully observed, $\lambda_k \in [0, 1]$ and i.i.d.

Audiovisual localization: Oracle dynamic stream weights

Assumption: $\boldsymbol{x}_k$, $\boldsymbol{y}_{\mathrm{A},k}$, $\boldsymbol{y}_{\mathrm{V},k}$, $k = 1, \ldots, K$ fully observed, $\lambda_k \in [0, 1]$ and i.i.d.

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

$$\Leftrightarrow \quad \log\{p(\boldsymbol{x}_k, \boldsymbol{y}_{\mathrm{A},k}, \boldsymbol{y}_{\mathrm{V},k}, \lambda_k)\} = \lambda_k \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} + (1 - \lambda_k) \log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} + c$$

Audiovisual localization: Oracle dynamic stream weights

Assumption: $\boldsymbol{x}_k,\ \boldsymbol{y}_{\mathrm{A},k},\ \boldsymbol{y}_{\mathrm{V},k},\ k = 1, \ldots, K$ fully observed, $\lambda_k \in [0,\ 1]$ and i.i.d.

$$p(\boldsymbol{x}_k,\ \boldsymbol{y}_{\mathrm{A},k},\ \boldsymbol{y}_{\mathrm{V},k},\ \lambda_k) \propto p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)^{\lambda_k} p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)^{1-\lambda_k}$$
$$\Leftrightarrow \quad \log\{p(\boldsymbol{x}_k,\ \boldsymbol{y}_{\mathrm{A},k},\ \boldsymbol{y}_{\mathrm{V},k},\ \lambda_k)\} = \lambda_k \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} + (1-\lambda_k)\log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} + c$$

Problem: Direct optimization not feasible.

Audiovisual localization: Oracle dynamic stream weights

Assumption: $\boldsymbol{x}_k$, $\boldsymbol{y}_{\mathrm{A},k}$, $\boldsymbol{y}_{\mathrm{V},k}$, $k = 1, \ldots, K$ fully observed, $\lambda_k \in [0, 1]$ and i.i.d.

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

$$\Leftrightarrow \quad \log\{p(\boldsymbol{x}_k, \boldsymbol{y}_{\mathrm{A},k}, \boldsymbol{y}_{\mathrm{V},k}, \lambda_k)\} = \lambda_k \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} + (1 - \lambda_k) \log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} + c$$

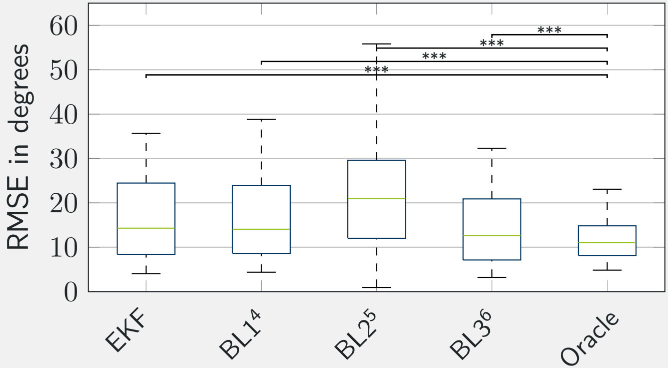
Problem: Direct optimization not feasible.

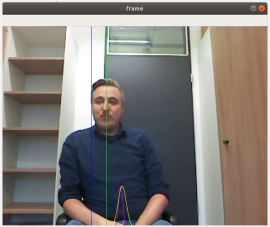
Solution: Impose prior on λ_k, e.g. Gaussian or symmetric Beta[3] distribution.

$$J(\lambda_k) = \lambda_k \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} + (1 - \lambda_k) \log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} + \log\{p(\lambda_k)\}$$

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

Audiovisual localization: Results I





[∗ ∗ ∗] $p < 0.001$

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

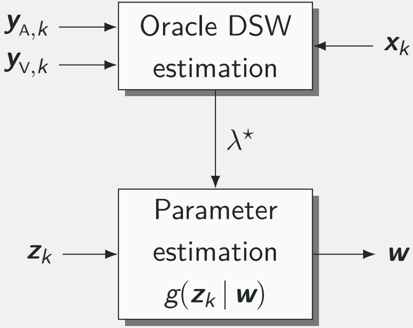
[5] S. Gerlach et al.: *2D audio-visual localization in home environments using a particle filter*, ITG Symp., 2012

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

Audiovisual localization: Learning dynamic stream weights

Supervised learning approach

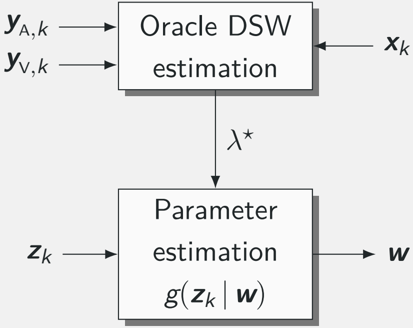
Oracle DSW serve as targets



Audiovisual localization: Learning dynamic stream weights

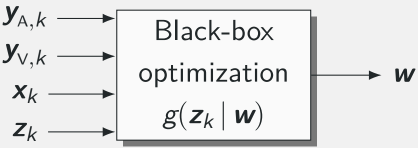
Supervised learning approach

Oracle DSW serve as targets



Evolutionary[7] learning approach

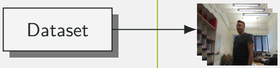
Direct optimization of localization error



[7] D. Wierstra et al.: *Natural evolution strategies*, Journal of machine learning research, vol. 15, 2014

Audiovisual localization: Learning dynamic stream weights

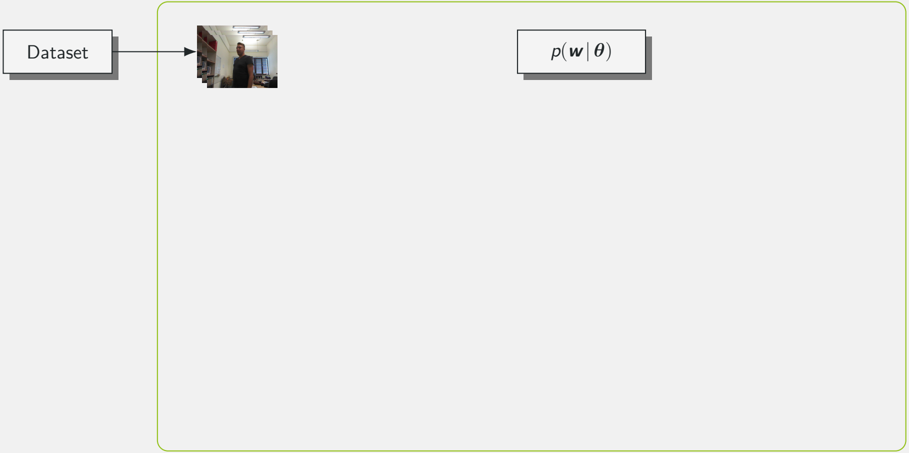
Training procedure[8]



[8] C. Schymura et al.: *Learning dynamic stream weights for linear dynamical systems using natural evolution strategies*, ICASSP, 2019

Audiovisual localization: Learning dynamic stream weights

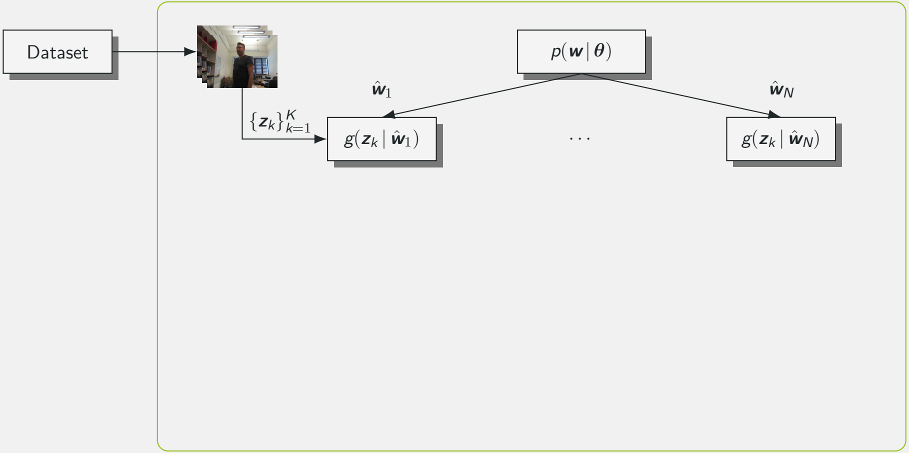
Training procedure[8]



[8] C. Schymura et al.: *Learning dynamic stream weights for linear dynamical systems using natural evolution strategies*, ICASSP, 2019

Audiovisual localization: Learning dynamic stream weights

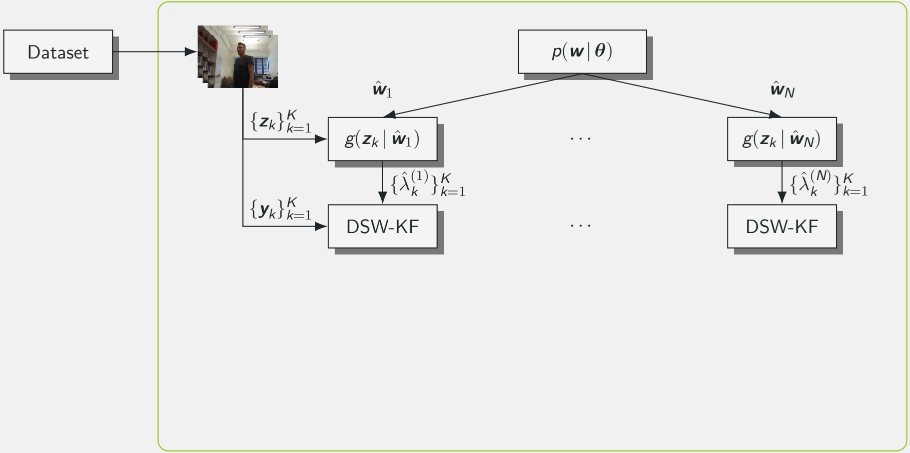
Training procedure[8]



[8] C. Schymura et al.: *Learning dynamic stream weights for linear dynamical systems using natural evolution strategies*, ICASSP, 2019

Audiovisual localization: Learning dynamic stream weights

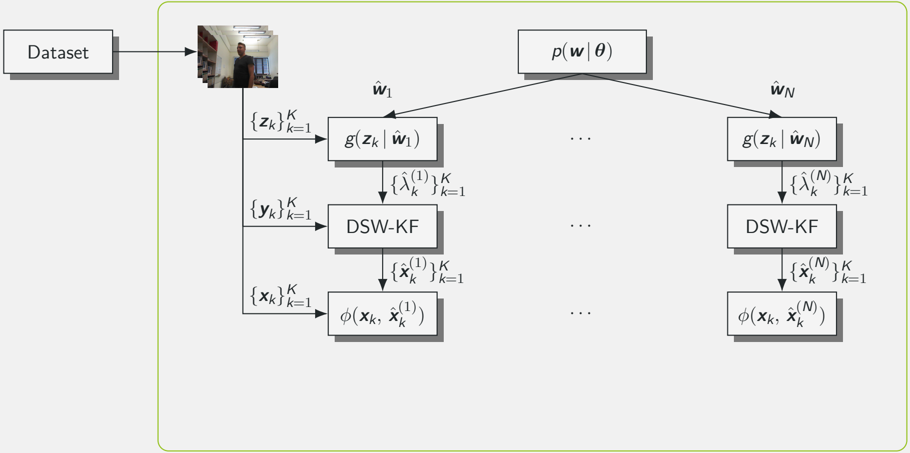
Training procedure[8]



[8] C. Schymura et al.: *Learning dynamic stream weights for linear dynamical systems using natural evolution strategies*, ICASSP, 2019

Audiovisual localization: Learning dynamic stream weights

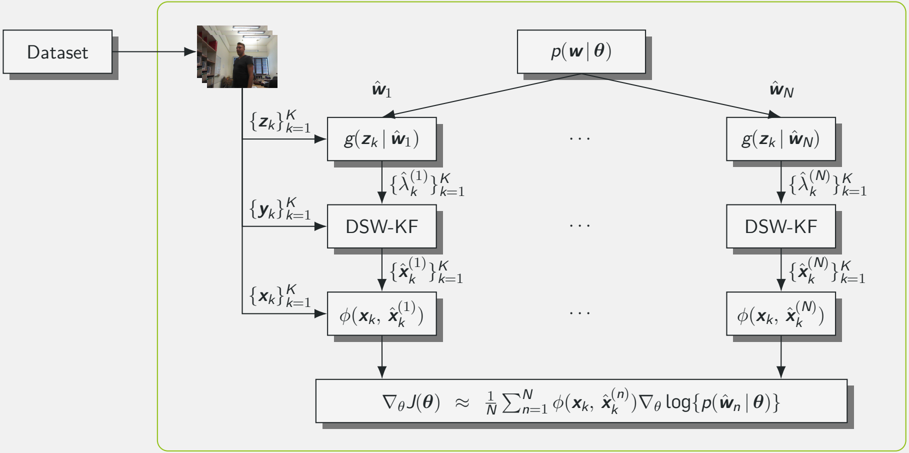
Training procedure[8]



[8] C. Schymura et al.: *Learning dynamic stream weights for linear dynamical systems using natural evolution strategies*, ICASSP, 2019

Audiovisual localization: Learning dynamic stream weights

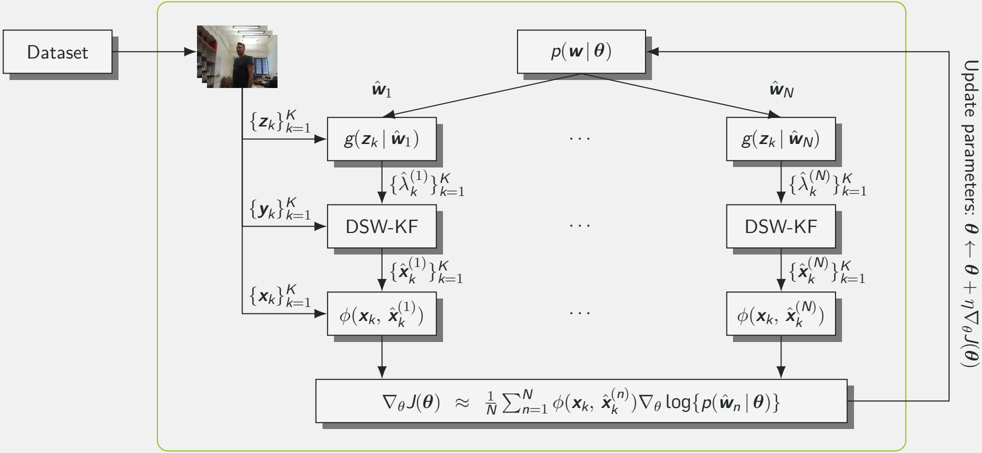
Training procedure[8]



[8] C. Schymura et al.: *Learning dynamic stream weights for linear dynamical systems using natural evolution strategies*, ICASSP, 2019

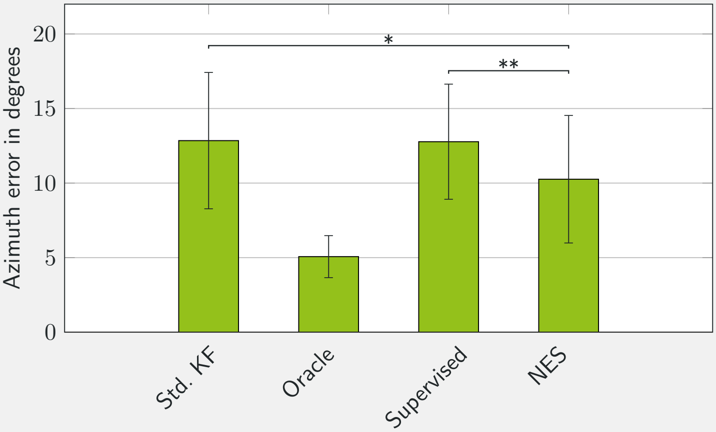
Audiovisual localization: Learning dynamic stream weights

Training procedure[8]



[8] C. Schymura et al.: *Learning dynamic stream weights for linear dynamical systems using natural evolution strategies*, ICASSP, 2019

Audiovisual localization: Results II

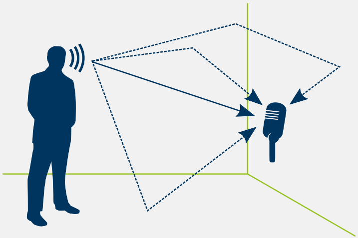


[$*$] $p < 0.05$
[$**$] $p < 0.01$

Part II

Causal models

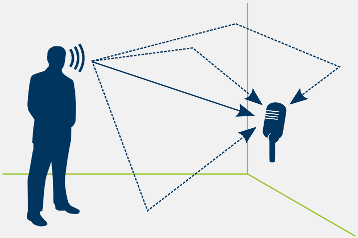
Causal models: Problem statement



Task: Sound source localization in reverberant rooms using spherical microphone arrays[9].

[9] O. Nadiri, B. Rafaely: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, IEEE Trans. on Audio, Speech, and Language Processing, vol. 22, 2014

Causal models: Problem statement

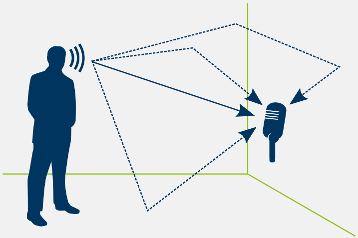


Task: Sound source localization in reverberant rooms using spherical microphone arrays[9].

- Direct-path dominance test-based direction-of-arrival estimation.

[9] O. Nadiri, B. Rafaely: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, IEEE Trans. on Audio, Speech, and Language Processing, vol. 22, 2014

Causal models: Problem statement

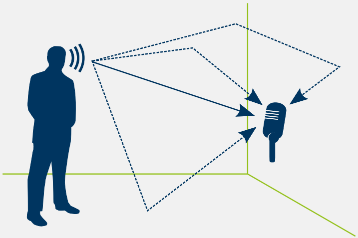


Task: Sound source localization in reverberant rooms using spherical microphone arrays[9].

- Direct-path dominance test-based direction-of-arrival estimation.
- Clustering of estimated DoAs using Gaussian mixture models.

[9] O. Nadiri, B. Rafaely: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, IEEE Trans. on Audio, Speech, and Language Processing, vol. 22, 2014

Causal models: Problem statement



Task: Sound source localization in reverberant rooms using spherical microphone arrays[9].

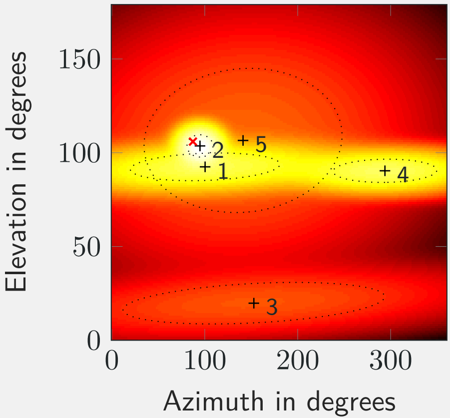
- Direct-path dominance test-based direction-of-arrival estimation.
- Clustering of estimated DoAs using Gaussian mixture models.
- Speaker DoA determined by selecting the dominant Gaussian component(s) of the Gaussian mixture model.

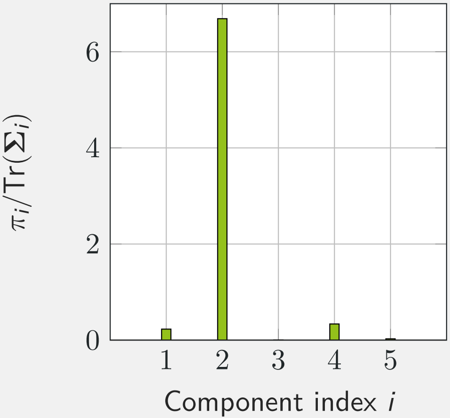
[9] O. Nadiri, B. Rafaely: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, IEEE Trans. on Audio, Speech, and Language Processing, vol. 22, 2014

Causal models: GMM-based DoA clustering

Reverberation time: $T_{60} = 0.5\,\mathrm{s}$

$\log p(\boldsymbol{\theta} \mid \{\pi_i,\, \boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i\}_{i=1}^{C})$

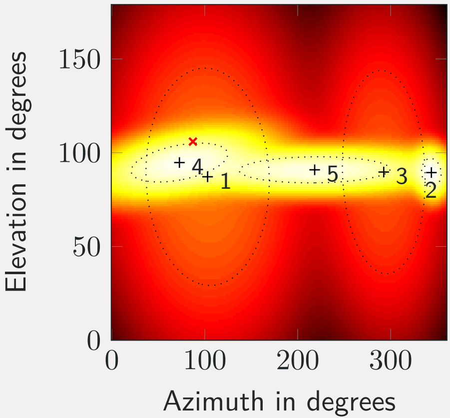


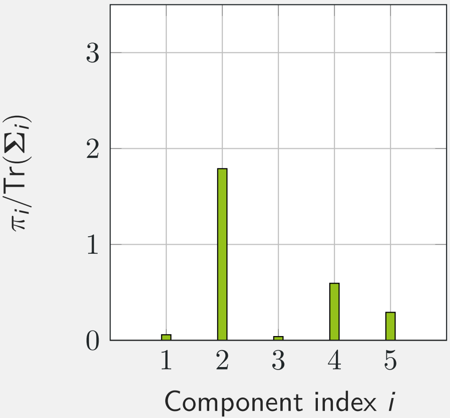


Causal models: GMM-based DoA clustering

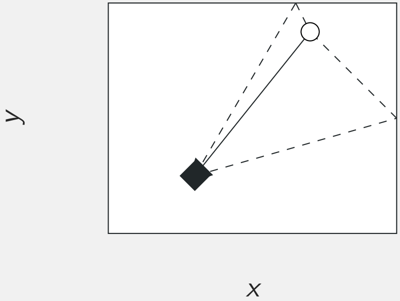
Reverberation time: $T_{60} = 2.0\,\mathrm{s}$

$$\log p(\boldsymbol{\theta} \mid \{\pi_i,\ \boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i\}_{i=1}^{C})$$

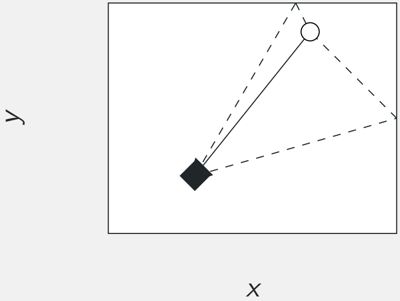


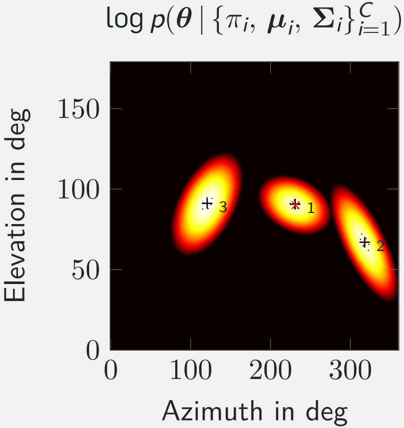


Causal models: Toy example

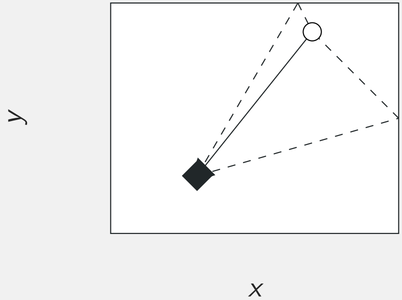


Causal models: Toy example

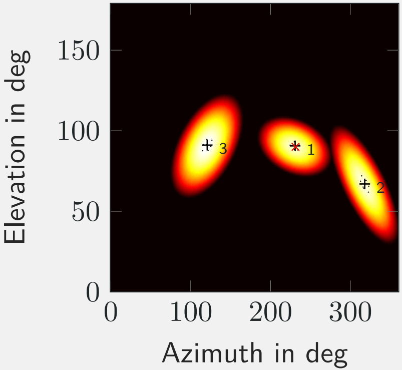




Causal models: Toy example

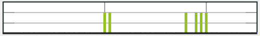


$\log p(\boldsymbol{\theta} \mid \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{C})$



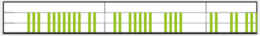
Causal models: Granger causality test[10]

x_k



k

y_k

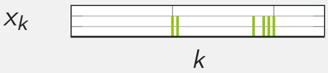


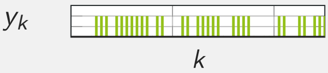
k

Granger causality: Does y_k **significantly contribute** to predicting x_k?

[10] C. W. J. Granger: *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica, vol. 37, 1969

Causal models: Granger causality test[10]





Granger causality: Does y_k **significantly contribute** to predicting x_k?

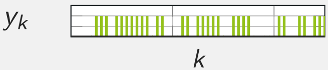
1. Fit autoregressive models

$$x_k = \sum_{\kappa=1}^{m} a_{xx,\kappa} x_{k-\kappa} + \sum_{\kappa=1}^{m} a_{xy,\kappa} y_{k-\kappa} + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0,\, \sigma^2)$$

$$x_k = \sum_{\kappa=1}^{m} \tilde{a}_{xx,\kappa} x_{k-\kappa} + \tilde{\epsilon}_k, \quad \tilde{\epsilon}_k \sim \mathcal{N}(0,\, \tilde{\sigma}^2)$$

[10] C. W. J. Granger: *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica, vol. 37, 1969

Causal models: Granger causality test[10]



Granger causality: Does y_k **significantly contribute** to predicting x_k?

1. Fit autoregressive models

$$x_k = \sum_{\kappa=1}^{m} a_{xx,\kappa} x_{k-\kappa} + \sum_{\kappa=1}^{m} a_{xy,\kappa} y_{k-\kappa} + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0,\, \sigma^2)$$

$$x_k = \sum_{\kappa=1}^{m} \tilde{a}_{xx,\kappa} x_{k-\kappa} + \tilde{\epsilon}_k, \quad \tilde{\epsilon}_k \sim \mathcal{N}(0,\, \tilde{\sigma}^2)$$

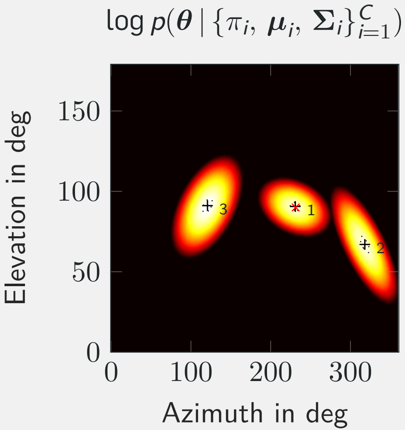
2. Evaluate null hypothesis $H_0 \ : \ a_{xy,\kappa} = 0 \ \forall \kappa$ via the F-test statistic

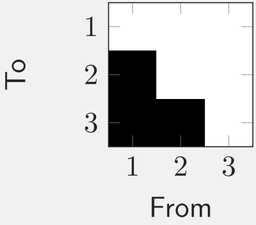
$$\mathcal{F}_{\boldsymbol{Y} \to \boldsymbol{X}} \equiv \frac{\tilde{\sigma}^2}{\sigma^2}$$

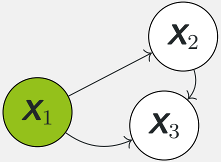
[10] C. W. J. Granger: *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica, vol. 37, 1969

Causal models: Causal graph and root node selection

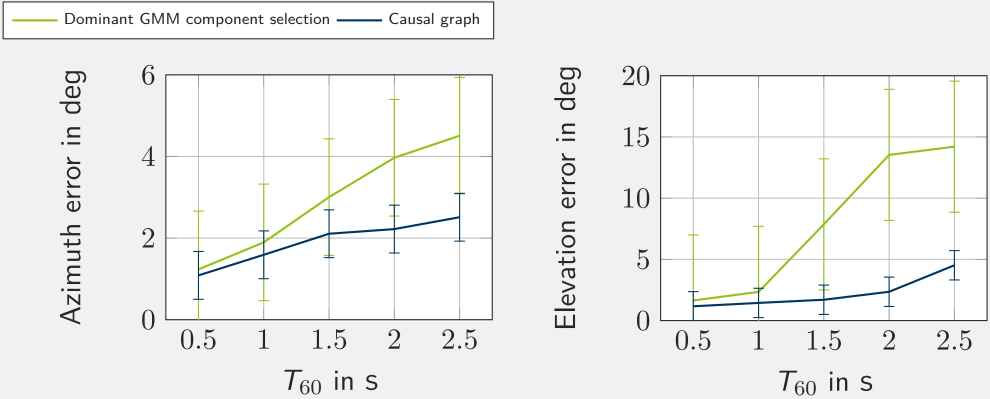
Constructing Granger matrix and causal graph via pair-wise Granger causality test:







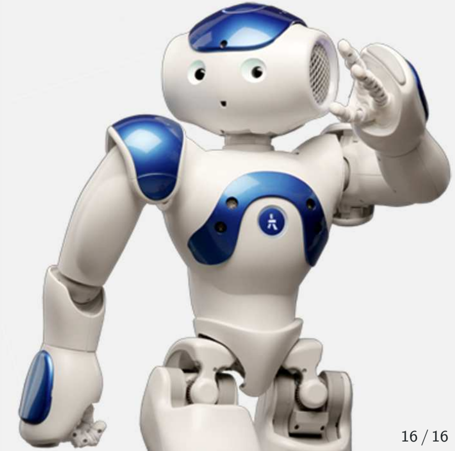
Causal models: Results



Conclusions

Conclusions

- Dynamic stream weights can benefit audiovisual speaker localization performance and provide an **additional notion of uncertainty**.

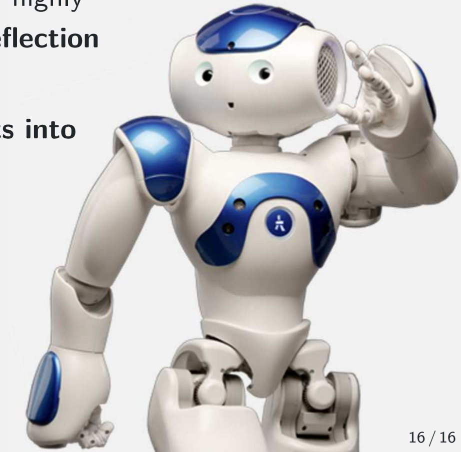


Conclusions

- Dynamic stream weights can benefit audiovisual speaker localization performance and provide an **additional notion of uncertainty**.
- Causal reasoning helps localizing sound sources in highly reverberant environments and allows to **depict reflection patterns via a causal graph**.

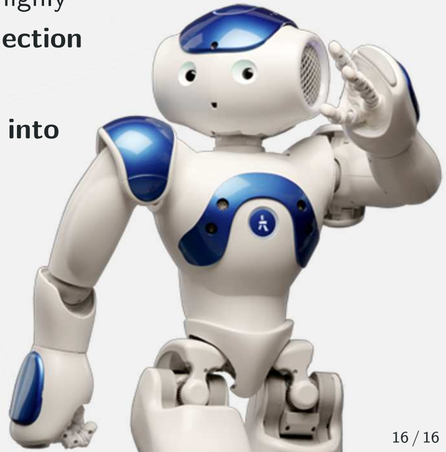
Conclusions

- Dynamic stream weights can benefit audiovisual speaker localization performance and provide an **additional notion of uncertainty**.
- Causal reasoning helps localizing sound sources in highly reverberant environments and allows to **depict reflection patterns via a causal graph**.
- Cognitive models in general yield valuable **insights into model behavior**.



Conclusions

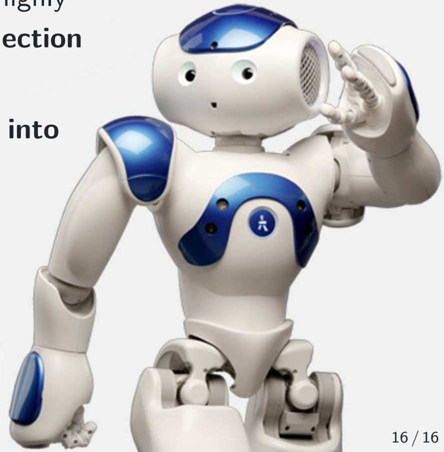
- Dynamic stream weights can benefit audiovisual speaker localization performance and provide an **additional notion of uncertainty**.
- Causal reasoning helps localizing sound sources in highly reverberant environments and allows to **depict reflection patterns via a causal graph**.
- Cognitive models in general yield valuable **insights into model behavior**.
- Promising direction for future research: How to integrate modern deep learning techniques without sacrificing model interpretability?



Conclusions

- Dynamic stream weights can benefit audiovisual speaker localization performance and provide an **additional notion of uncertainty**.
- Causal reasoning helps localizing sound sources in highly reverberant environments and allows to **depict reflection patterns via a causal graph**.
- Cognitive models in general yield valuable **insights into model behavior**.
- Promising direction for future research: How to integrate modern deep learning techniques without sacrificing model interpretability?

Thank you for your attention!



Supplementary material

DSW-EKF: Derivation I

Prediction step

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

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

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

Prediction step (identical to standard EKF)

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

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

DSW-EKF: Derivation II

Update step

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

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

Update step

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

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

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

DSW-EKF: Inference

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

can be expressed as

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

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

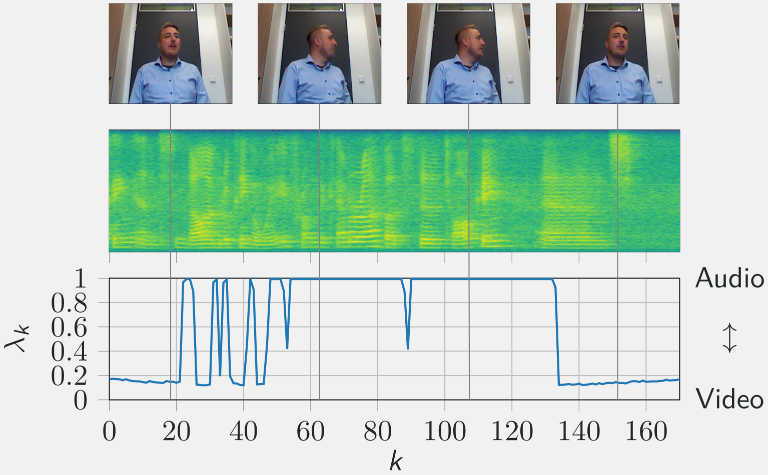
Modified Kalman gain computation using the binomial inverse theorem[11]

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

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

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

Audiovisual localization: Dynamic stream weights



ODSW estimation: Gaussian prior

$$J(\lambda_k) = \lambda_k \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} + (1-\lambda_k)\log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} + \log\{p(\lambda_k)\}$$

with

$$\log\{p(\lambda_k)\} = -\frac{1}{2}\frac{(\lambda_k-\mu_\lambda)^2}{\sigma_\lambda^2} + \text{const.}$$

yields

$$\frac{dJ(\lambda_k)}{d\lambda_k} = \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} - \log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} - \frac{1}{\sigma_\lambda^2}(\lambda_k-\mu_\lambda)$$

$$\Rightarrow \quad \lambda_k^\star = \mu_\lambda + \sigma_\lambda^2 \log\left\{\frac{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)}{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)}\right\}$$

ODSW estimation: Symmetric Beta prior I

$$J(\lambda_k) = \lambda_k \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} + (1-\lambda_k)\log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} + \log\{p(\lambda_k)\}$$

with

$$p(\lambda_k) = \frac{1}{\mathrm{B}(\alpha_\lambda,\, \alpha_\lambda)} \lambda_k^{\alpha_\lambda - 1}(1-\lambda_k)^{\alpha_\lambda - 1}$$

yields

$$J_{\mathrm{Beta}}(\lambda_k) = \lambda_k \log\{p(\boldsymbol{y}_{\mathrm{A},k}|\boldsymbol{x}_k)\} + (1-\lambda_k)\log\{p(\boldsymbol{y}_{\mathrm{V},k}|\boldsymbol{x}_k)\} \\ + (\alpha_\lambda - 1)\Big(\log\{\lambda_k\} + \log\{1-\lambda_k\}\Big) + \text{const.}$$

$$\Rightarrow \quad \lambda_k^\star = \max_{\lambda_k} J_{\mathrm{Beta}}(\lambda_k) \qquad \text{s. t.} \quad 0 < \lambda_k < 1$$

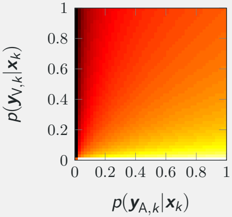
ODSW estimation: Symmetric Beta prior II

$J_{\text{Beta}}(\lambda_k)$ is a concave function:

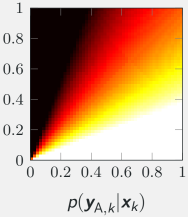
$$\frac{dJ_{\text{Beta}}(\lambda_k)}{d\,\lambda_k} = \log\left\{\frac{p(\boldsymbol{y}_{\text{A},k}|\boldsymbol{x}_k)}{p(\boldsymbol{y}_{\text{V},k}|\boldsymbol{x}_k)}\right\} + (\alpha_\lambda - 1)\left(\frac{1}{\lambda_k} + \frac{1}{\lambda_k - 1}\right)$$

$$\frac{d^2 J_{\text{Beta}}(\lambda_k)}{d\,\lambda_k^2} = (\alpha_\lambda - 1)\left(\frac{1}{\lambda_k^2} + \frac{1}{(\lambda_k - 1)^2}\right) < 0 \;\forall\, \alpha_\lambda > 1$$

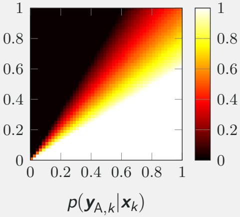
ODSW examples



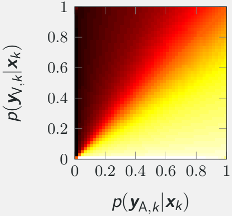
(a) $\mu_\lambda = 0.5,\ \sigma_\lambda^2 = 0.1$



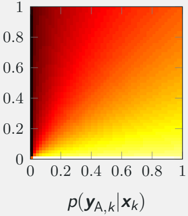
(b) $\mu_\lambda = 0.5,\ \sigma_\lambda^2 = 0.5$



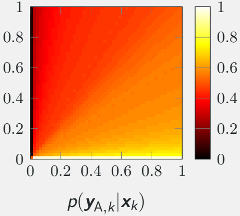
(c) $\mu_\lambda = 0.5,\ \sigma_\lambda^2 = 1.0$



(d) $\alpha_\lambda = 1.1$



(e) $\alpha_\lambda = 1.5$



(f) $\alpha_\lambda = 2.5$

Audiovisual localization: Experimental setup

- Three audiovisual datasets.
- Acoustic front-end: SRP-PHAT
- Visual front-end: YOLOFace[12]
- Constant velocity linear dynamics model and nonlinear rotating vector observation models.
- Leave-one-out cross-validation paradigm.

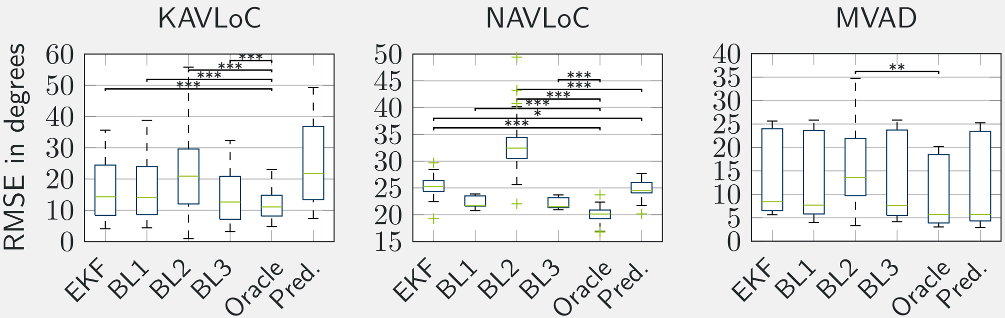
[12] J. Redmon, A. Farhadi: *YOLOv3: An Incremental Improvement*, arXiv, 2018

Audiovisual localization: Results (contd.)

Tabelle 1: Root mean squared errors in degrees.

	Undistorted	Signal-to-noise ratio			Image rotation		
		0 dB	15 dB	30 dB	10°	20°	30°
KAVLoC ($N = 70$)							
EKF (Audio)	11.92 ± 2.93	20.22 ± 4.86	14.24 ± 3.72	11.95 ± 2.93	11.92 ± 2.93	11.92 ± 2.93	11.92 ± 2.93
EKF (Video)	6.78 ± 3.16	6.78 ± 3.16	6.78 ± 3.16	6.78 ± 3.16	6.94 ± 3.37	7.88 ± 3.95	9.32 ± 4.76
EKF (Audiovisual)	8.64 ± 2.43	11.27 ± 2.57	9.04 ± 2.40	8.64 ± 2.53	7.48 ± 2.70	7.77 ± 2.45	8.05 ± 2.46
ODSW-EKF (Gaussian)	$5.87 \pm 2.79^\star$	$5.87 \pm 2.77^\star$	$5.85 \pm 2.79^\star$	$5.87 \pm 2.79^\star$	$6.09 \pm 2.99^\star$	6.99 ± 3.47	7.95 ± 3.77
ODSW-EKF (Beta)	$5.85 \pm 2.86^\star$	$5.84 \pm 2.79^\star$	$5.79 \pm 2.85^\star$	$5.87 \pm 2.87^\star$	$6.06 \pm 3.04^\star$	6.90 ± 3.49	7.86 ± 3.80
NAVLoC ($N = 400$)							
EKF (Audio)	21.55 ± 0.53	21.60 ± 0.54	21.59 ± 0.54	21.59 ± 0.54	21.55 ± 0.53	21.55 ± 0.53	21.55 ± 0.53
EKF (Video)	19.00 ± 0.37	19.00 ± 0.37	19.00 ± 0.37	19.00 ± 0.37	19.15 ± 0.36	19.47 ± 0.34	20.20 ± 0.72
EKF (Audiovisual)	21.36 ± 0.15	21.37 ± 0.16	21.37 ± 0.15	21.37 ± 0.16	21.42 ± 0.15	21.52 ± 0.15	21.72 ± 0.19
ODSW-EKF (Gaussian)	$15.69 \pm 0.57^\star$	$15.69 \pm 0.64^\star$	$15.69 \pm 0.64^\star$	$15.69 \pm 0.64^\star$	$16.08 \pm 0.63^\star$	$16.84 \pm 0.57^\star$	$18.20 \pm 0.98^\star$
ODSW-EKF (Beta)	$15.69 \pm 0.57^\star$	$15.69 \pm 0.64^\star$	$15.69 \pm 0.64^\star$	$15.69 \pm 0.64^\star$	$16.08 \pm 0.63^\star$	$16.84 \pm 0.57^\star$	$18.21 \pm 0.98^\star$
MVAD ($N = 6$)							
EKF (Audio)	15.18 ± 8.62	20.47 ± 11.96	19.61 ± 12.13	15.53 ± 7.45	15.18 ± 8.62	15.18 ± 8.62	15.18 ± 8.62
EKF (Video)	10.07 ± 9.51	10.07 ± 9.51	10.07 ± 9.51	10.07 ± 9.51	10.61 ± 9.33	11.43 ± 9.24	12.36 ± 8.99
EKF (Audiovisual)	12.77 ± 8.82	13.57 ± 10.55	13.87 ± 9.67	13.13 ± 8.27	12.90 ± 8.50	13.23 ± 8.70	13.68 ± 8.73
ODSW-EKF (Gaussian)	8.85 ± 7.37	10.12 ± 9.30	9.39 ± 8.16	8.65 ± 7.04	8.91 ± 6.97	9.80 ± 6.81	10.67 ± 6.55
ODSW-EKF (Beta)	8.86 ± 7.37	10.12 ± 9.30	9.39 ± 8.16	8.66 ± 7.04	8.90 ± 6.96	9.80 ± 6.81	10.67 ± 6.55

Audiovisual localization: Results (contd.)

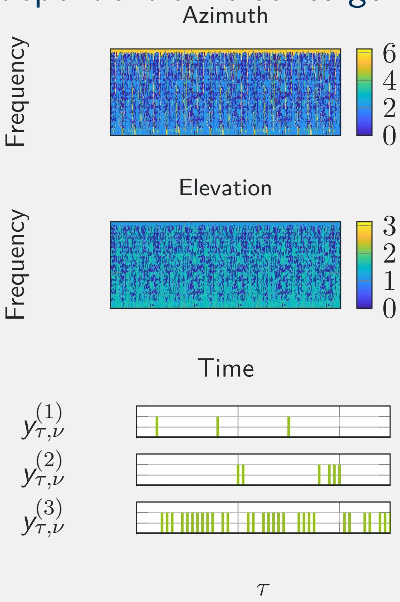


NES: Gradient approximation

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\{f(\boldsymbol{w})\} = \int f(\boldsymbol{w}) p(\boldsymbol{w} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{w}.$$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \int f(\boldsymbol{w}) p(\boldsymbol{w} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{w} \\
&= \int f(\boldsymbol{w}) \nabla_{\boldsymbol{\theta}} p(\boldsymbol{w} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{w} \\
&= \int f(\boldsymbol{w}) \nabla_{\boldsymbol{\theta}} p(\boldsymbol{w} \,|\, \boldsymbol{\theta}) \frac{p(\boldsymbol{w} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{w} \,|\, \boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{w} \\
&= \int \Big(f(\boldsymbol{w}) \nabla_{\boldsymbol{\theta}} \log\{p(\boldsymbol{w} \,|\, \boldsymbol{\theta})\} \Big) p(\boldsymbol{w} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{w} \\
&= \mathbb{E}_{\boldsymbol{\theta}} \Big\{ f(\boldsymbol{w}) \nabla_{\boldsymbol{\theta}} \log\{p(\boldsymbol{w} \,|\, \boldsymbol{\theta})\} \Big\} \approx \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{w}_m) \nabla_{\boldsymbol{\theta}} \log\{p(\boldsymbol{w}_m \,|\, \boldsymbol{\theta})\}
\end{aligned}$$

DoA-dependent time-series generation for GCT



1. Generate DoA time-series for each frequency bin:

$$\boldsymbol{\Theta}_\nu = \Big\{ \underbrace{\begin{bmatrix} \phi_{\tau,\nu} & \psi_{\tau,\nu} \end{bmatrix}^\mathsf{T}}_{\boldsymbol{\theta}^\mathsf{T}_{\tau,\nu}} \Big\}_{\tau=1}^{T}$$

2. Evaluate component-wise posteriors:

$$y^{(i)}_{\tau,\nu} = p(\boldsymbol{\theta}_{\tau,\nu} \,|\, \boldsymbol{\mu}_i, \, \boldsymbol{\Sigma}_i)$$

3. Generate time-series from posteriors:

$$\boldsymbol{y}_\nu = \{y^{(i)}_{\tau,\nu}\}_{\tau=1}^{T}$$