Learning Dynamic Stream Weights for Linear Dynamical Systems using Natural Evolution Strategies

ICASSP 2019

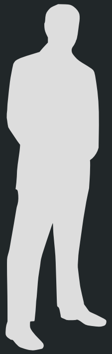
Christopher Schymura and Dorothea Kolossa

May 16th, 2019

RUHR
UNIVERSITÄT
BOCHUM

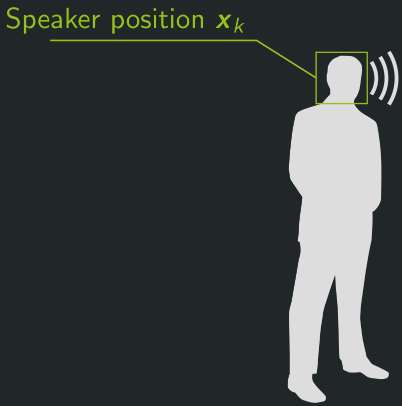
RUB

Audiovisual speaker tracking

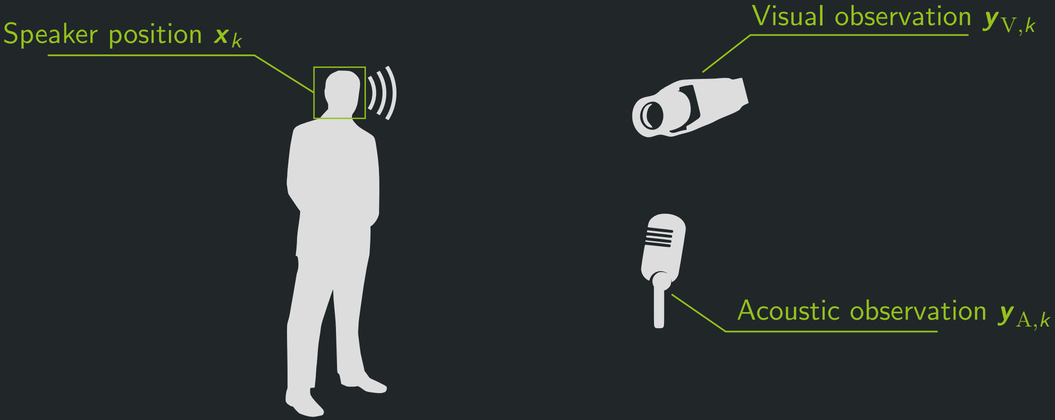


Audiovisual speaker tracking

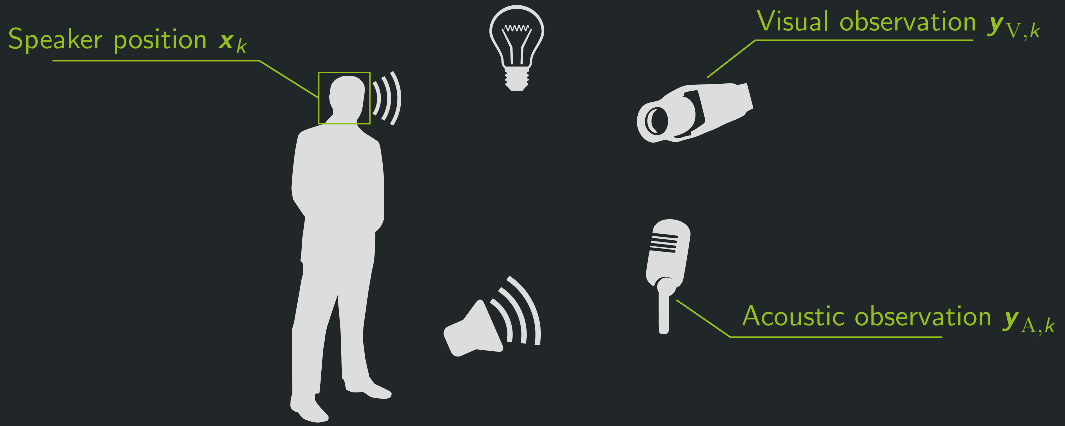
Audiovisual speaker tracking



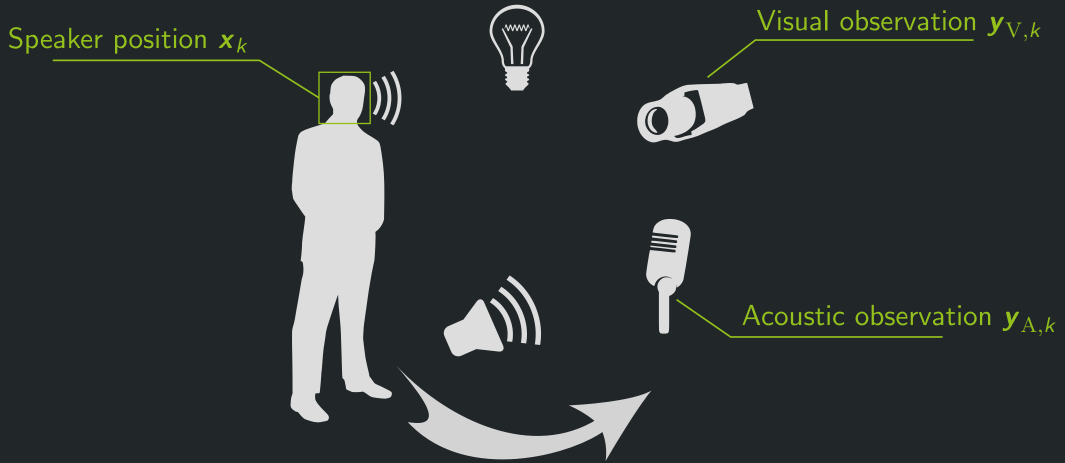
Audiovisual speaker tracking



Audiovisual speaker tracking



Audiovisual speaker tracking

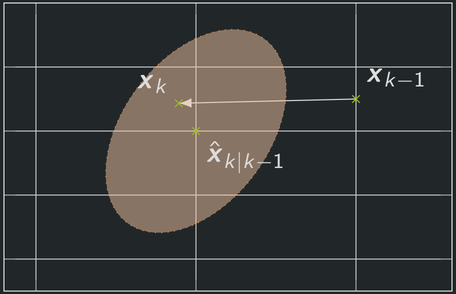


Audiovisual speaker tracking

Prediction step

System dynamics:

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

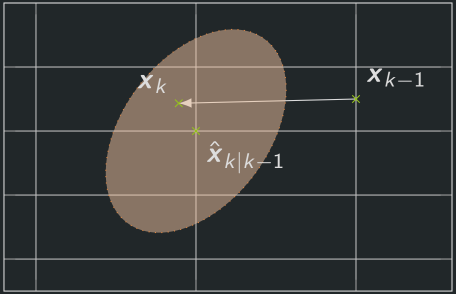


Audiovisual speaker tracking

Prediction step

System dynamics:

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



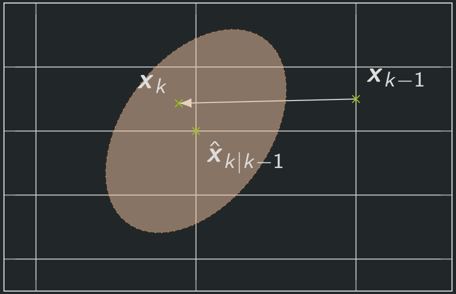
$$p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1}) = \int p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1})\ p(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{\mathrm{A},k-1},\ \boldsymbol{Y}_{\mathrm{V},k-1})\ d\boldsymbol{x}_{k-1}$$

Audiovisual speaker tracking

Prediction step

System dynamics:

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



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

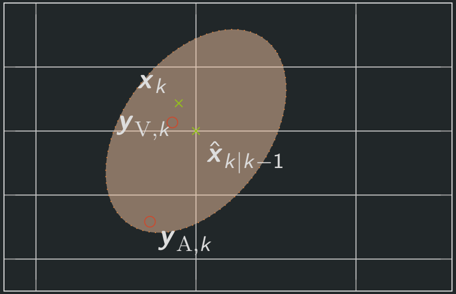
Audiovisual speaker tracking

Observation

Observation model:

$$\boldsymbol{y}_k = \begin{bmatrix} \boldsymbol{y}_{\mathrm{A},k} & \boldsymbol{y}_{\mathrm{V},k} \end{bmatrix}^{\mathrm{T}} = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{w}_k$$

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



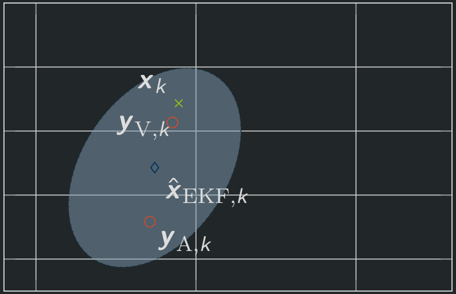
Audiovisual speaker tracking

Update step (standard Kalman filter)

Observation model:

$$\boldsymbol{y}_k = \begin{bmatrix} \boldsymbol{y}_{\mathrm{A},k} & \boldsymbol{y}_{\mathrm{V},k} \end{bmatrix}^{\mathrm{T}} = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{w}_k$$

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



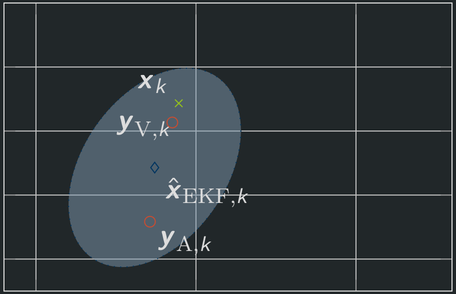
Audiovisual speaker tracking

Update step (standard Kalman filter)

Observation model:

$$\boldsymbol{y}_k = \begin{bmatrix} \boldsymbol{y}_{\mathrm{A},k} & \boldsymbol{y}_{\mathrm{V},k} \end{bmatrix}^{\mathrm{T}} = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{w}_k$$

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



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

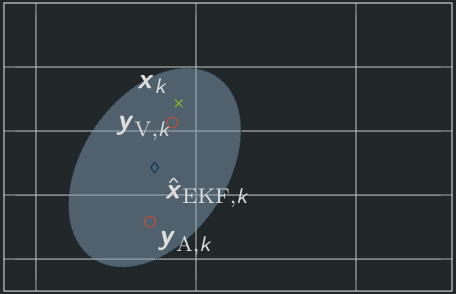
Audiovisual speaker tracking

Update step (standard Kalman filter)

Observation model:

$$\boldsymbol{y}_k = \begin{bmatrix} \boldsymbol{y}_{\mathrm{A},k} & \boldsymbol{y}_{\mathrm{V},k} \end{bmatrix}^{\mathrm{T}} = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{w}_k$$

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



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

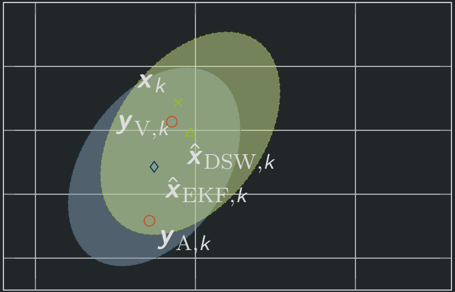
Audiovisual speaker tracking

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

Observation model:

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

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



[1] C. Schymura, T. Isenberg, D. Kolossa: *Extending Linear Dynamical Systems with Dynamic Stream Weights for Audiovisual Speaker*

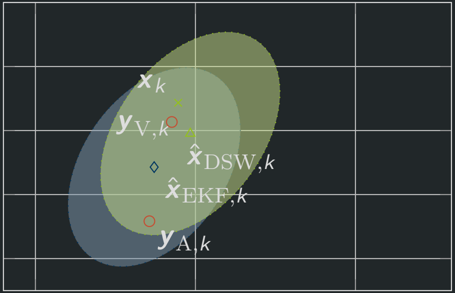
Audiovisual speaker tracking

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

Observation model:

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

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

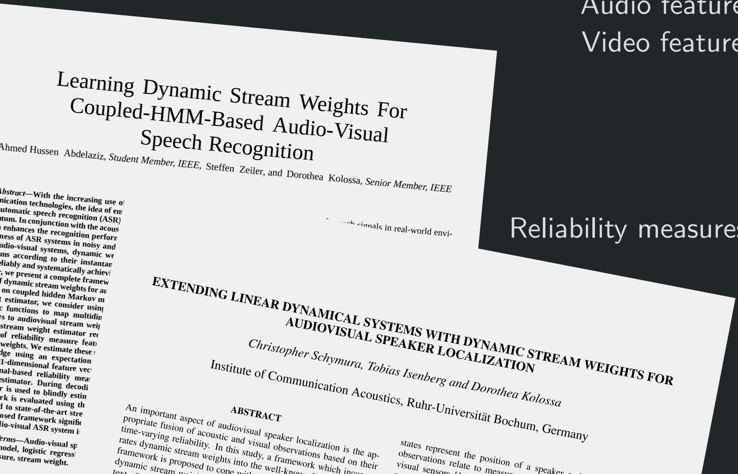


$$p(\boldsymbol{x}_k \,|\, \boldsymbol{Y}_{\mathrm{A},k}, \boldsymbol{Y}_{\mathrm{V},k}) \propto p(\boldsymbol{x}_k \,|\, \boldsymbol{Y}_{\mathrm{A},k-1}, \boldsymbol{Y}_{\mathrm{V},k-1}) \underbrace{p(\boldsymbol{y}_{\mathrm{A},k} \,|\, \boldsymbol{x}_k)^{\lambda_k}}_{\text{Acoustic model}} \underbrace{p(\boldsymbol{y}_{\mathrm{V},k}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

[1] C. Schymura, T. Isenberg, D. Kolossa: *Extending Linear Dynamical Systems with Dynamic Stream Weights for Audiovisual Speaker*

Learning dynamic stream weights

Standard approach: Supervised training with oracle dynamic stream weights


Learning Dynamic Stream Weights For Coupled-HMM-Based Audio-Visual Speech Recognition

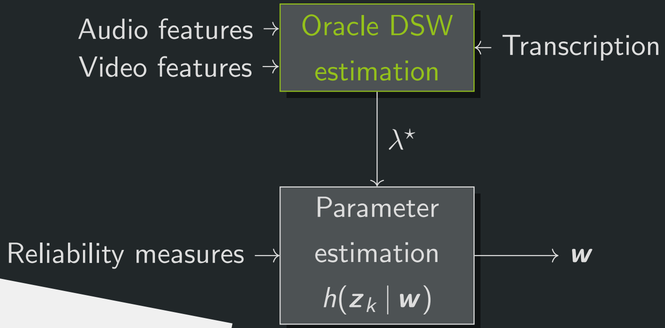
Ahmed Hussen Abdelaziz, *Student Member, IEEE,* Steffen Zeiler, and Dorothea Kolossa, *Senior Member, IEEE*

EXTENDING LINEAR DYNAMICAL SYSTEMS WITH DYNAMIC STREAM WEIGHTS FOR AUDIOVISUAL SPEAKER LOCALIZATION

Christopher Schymura, Tobias Isenberg and Dorothea Kolossa

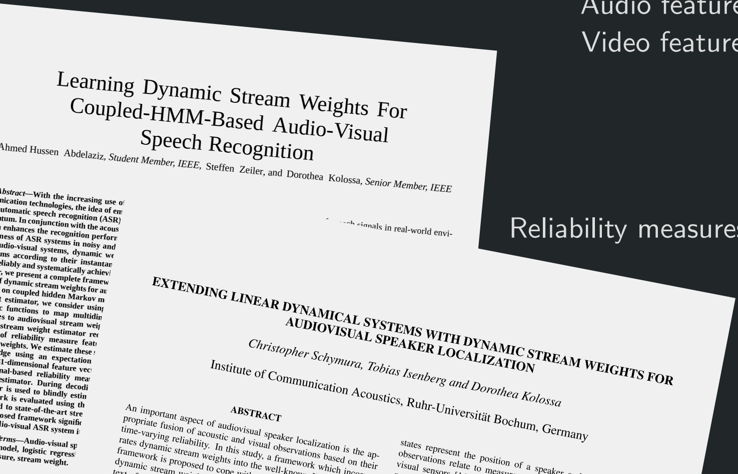
Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany

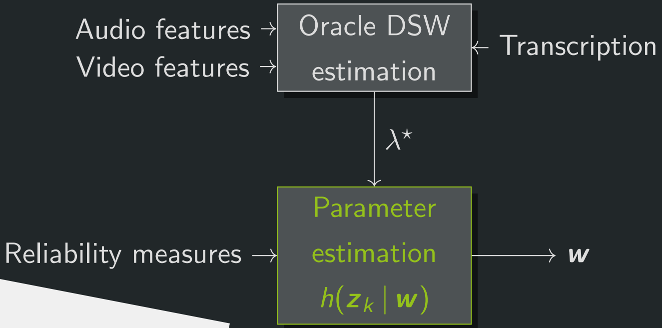
ABSTRACT



Learning dynamic stream weights

Standard approach: Supervised training with oracle dynamic stream weights





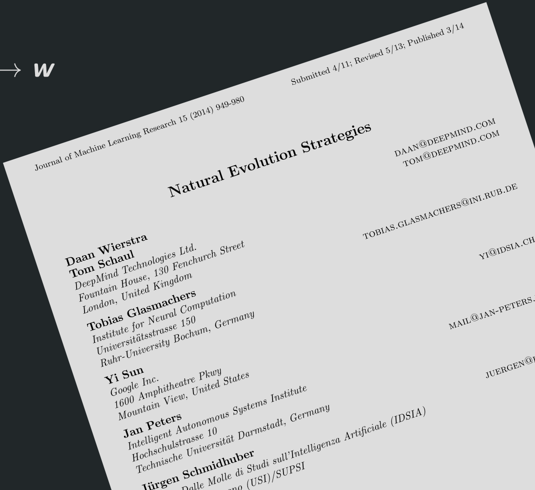
Learning dynamic stream weights

Proposed approach: Training with natural evolution strategies

Audio features →
Video features →
Speaker positions →
Realiability measures →

Black-box optimization $h(\boldsymbol{z}_k \mid \boldsymbol{w})$ → $\boldsymbol{w}$

- No oracle information required.
- Flexible choice of loss/fitness function.


Journal of Machine Learning Research 15 (2014) 949-980
Submitted 4/11; Revised 5/13; Published 3/14

Natural Evolution Strategies

Daan Wierstra DAAN@DEEPMIND.COM
Tom Schaul TOM@DEEPMIND.COM
DeepMind Technologies Ltd.
Fountain House, 130 Fenchurch Street
London, United Kingdom

Tobias Glasmachers TOBIAS.GLASMACHERS@INI.RUB.DE
Institute for Neural Computation
Universitätsstrasse 150
Ruhr-University Bochum, Germany

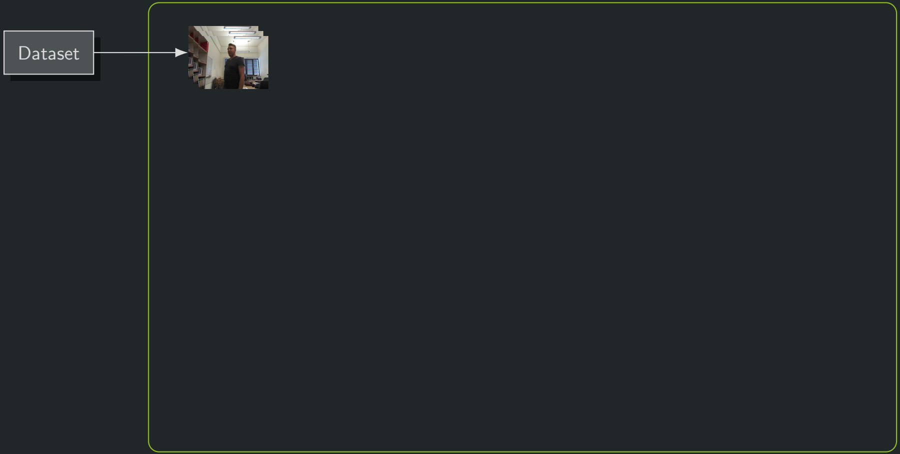
Yi Sun YI@IDSIA.CH
Google Inc.
1600 Amphitheatre Pkwy
Mountain View, United States

Jan Peters MAIL@JAN-PETERS.
Intelligent Autonomous Systems Institute
Hochschulstrasse 10
Technische Universität Darmstadt, Germany

Jürgen Schmidhuber JUERGEN@
Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
(USI)/SUPSI

Learning dynamic stream weights

Training procedure



Learning dynamic stream weights

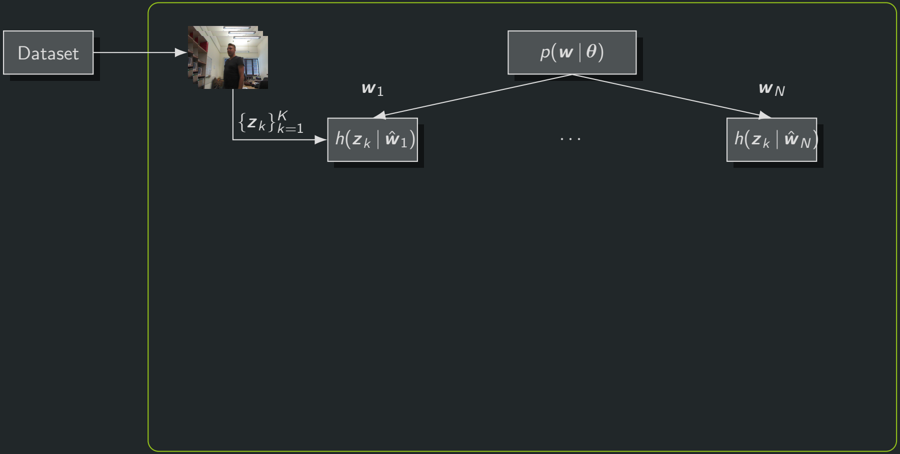
Training procedure

Dataset → 

$p(\boldsymbol{w} \mid \boldsymbol{\theta})$

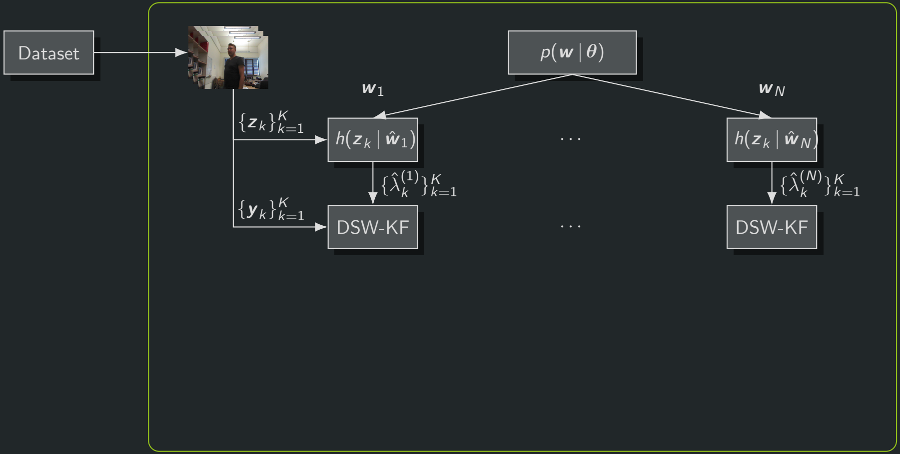
Learning dynamic stream weights

Training procedure



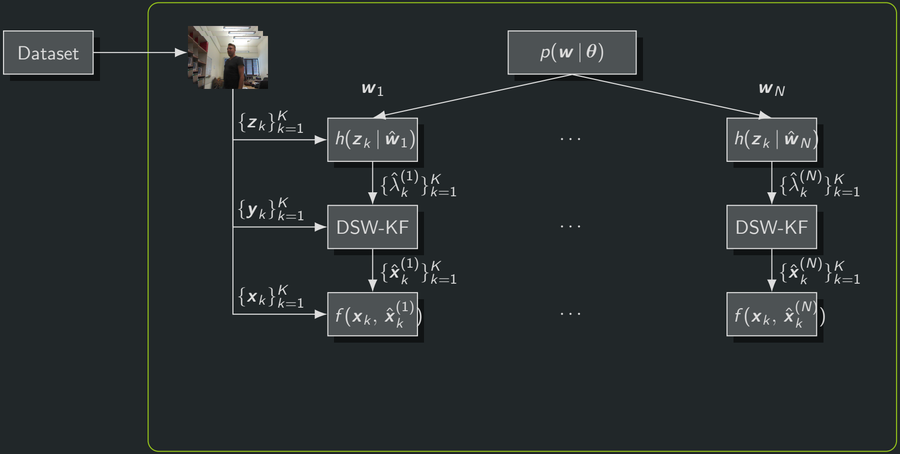
Learning dynamic stream weights

Training procedure



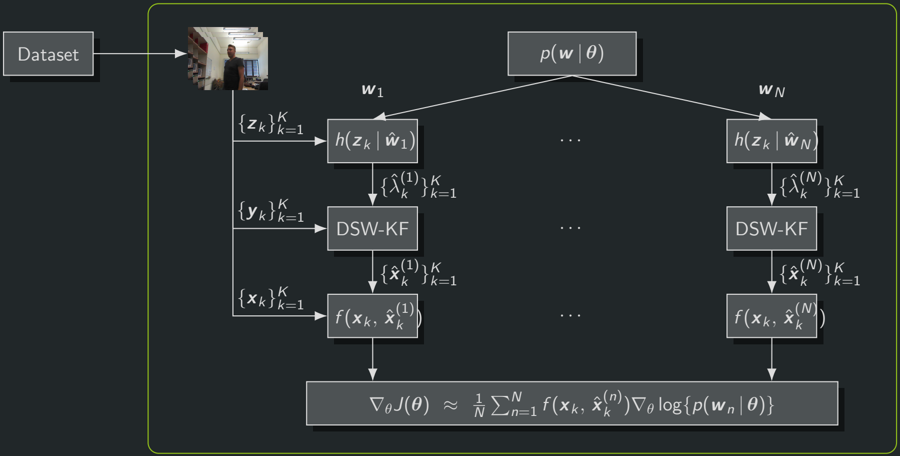
Learning dynamic stream weights

Training procedure



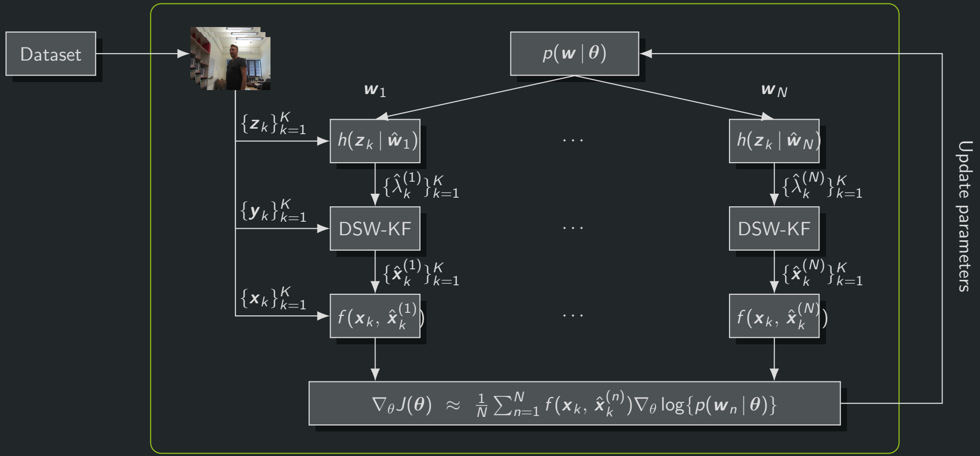
Learning dynamic stream weights

Training procedure



Learning dynamic stream weights

Training procedure



Learning dynamic stream weights

Implementation

- Reliability measures: instantaneous estimated a-priori SNR, acoustic and visual observation log-likelihoods[2].

[2] A. H. Abdelaziz, S. Zeiler, D. Kolossa: *Learning Dynamic Stream Weights for Coupled-HMM-Based Audio-Visual Speech Recognition*, 2015

Learning dynamic stream weights

Implementation

- Reliability measures: instantaneous estimated a-priori SNR, acoustic and visual observation log-likelihoods[2].
- Evaluation of two different DSW prediction models: logistic function and fully-connected feed-forward neural network.

[2] A. H. Abdelaziz, S. Zeiler, D. Kolossa: *Learning Dynamic Stream Weights for Coupled-HMM-Based Audio-Visual Speech Recognition*, 2015

Learning dynamic stream weights

Implementation

- Reliability measures: instantaneous estimated a-priori SNR, acoustic and visual observation log-likelihoods[2].
- Evaluation of two different DSW prediction models: logistic function and fully-connected feed-forward neural network.
- Separable natural evolution strategies (sNES) as optimizer:
  $p(\boldsymbol{w} \mid \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{\mu}_{\boldsymbol{w}}, \operatorname{diag}(\boldsymbol{\sigma}_{\boldsymbol{w}})\right)$

[2] A. H. Abdelaziz, S. Zeiler, D. Kolossa: *Learning Dynamic Stream Weights for Coupled-HMM-Based Audio-Visual Speech Recognition*, 2015

Learning dynamic stream weights

Implementation

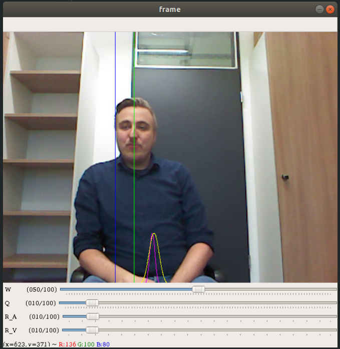
- Reliability measures: instantaneous estimated a-priori SNR, acoustic and visual observation log-likelihoods[2].
- Evaluation of two different DSW prediction models: logistic function and fully-connected feed-forward neural network.
- Separable natural evolution strategies (sNES) as optimizer:
  $p(\boldsymbol{w} \mid \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{\mu}_{\boldsymbol{w}}, \operatorname{diag}(\boldsymbol{\sigma}_{\boldsymbol{w}})\right)$
- Fitness function allowing direct optimization of instantaneous localization error:
  $f(\boldsymbol{w}) = -\frac{1}{M}\sum_{m=1}^{M}\frac{1}{K_m}\sum_{k=1}^{K_m}\left(\phi_k^{(m)} - \hat{\phi}_k^{(m)}(\boldsymbol{w})\right)^2$

[2] A. H. Abdelaziz, S. Zeiler, D. Kolossa: *Learning Dynamic Stream Weights for Coupled-HMM-Based Audio-Visual Speech Recognition*, 2015

Evaluation

Experimental setup

- Front-end: DPD-MUSIC[3] for acoustic localization, Viola-Jones[4] algorithm for visual localization.



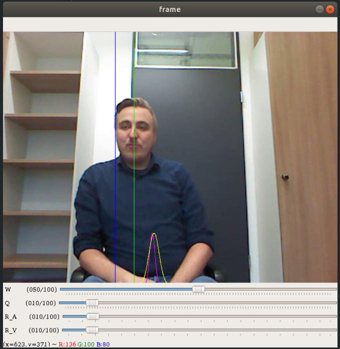
[3] Nadiri et al.: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, 2014

[4] P. Viola, M. Jones: *Rapid object detection using a boosted cascade of simple features*, 2001

Evaluation

Experimental setup

- Front-end: DPD-MUSIC[3] for acoustic localization, Viola-Jones[4] algorithm for visual localization.
- Dataset of audiovisual recordings in an office environment ($T_{60} \approx 350\,\mathrm{ms}$) using the Kinect.



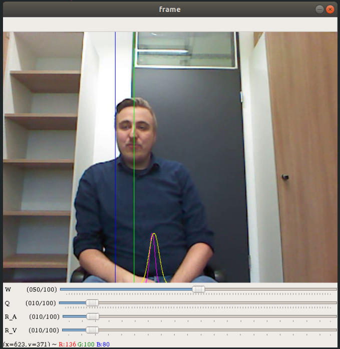
[3] Nadiri et al.: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, 2014

[4] P. Viola, M. Jones: *Rapid object detection using a boosted cascade of simple features*, 2001

Evaluation

Experimental setup

- Front-end: DPD-MUSIC[3] for acoustic localization, Viola-Jones[4] algorithm for visual localization.
- Dataset of audiovisual recordings in an office environment ($T_{60} \approx 350\,\mathrm{ms}$) using the Kinect.
- Constant velocity dynamics model.



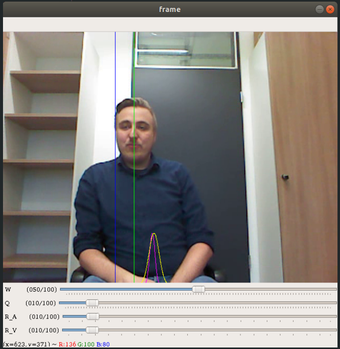
[3] Nadiri et al.: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, 2014

[4] P. Viola, M. Jones: *Rapid object detection using a boosted cascade of simple features*, 2001

Evaluation

Experimental setup

- Front-end: DPD-MUSIC[3] for acoustic localization, Viola-Jones[4] algorithm for visual localization.
- Dataset of audiovisual recordings in an office environment ($T_{60} \approx 350\,\mathrm{ms}$) using the Kinect.
- Constant velocity dynamics model.
- Baseline: Stream weight prediction models trained on oracle DSWs with SGD (same architecture)

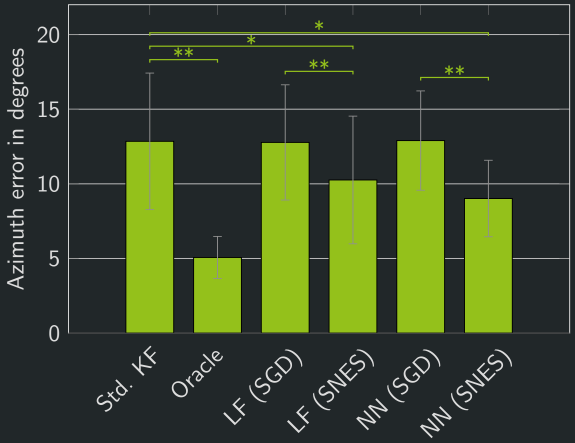


[3] Nadiri et al.: *Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test*, 2014

[4] P. Viola, M. Jones: *Rapid object detection using a boosted cascade of simple features*, 2001

Evaluation

Results



Statistical significance: * for $p < 0.05$ and ** for $p < 0.01$

Conclusions and outlook

- A DSW-based audiovisual speaker tracking system can benefit from black-box optimization approaches like NES (no oracle DSWs required).

Conclusions and outlook

- A DSW-based audiovisual speaker tracking system can benefit from black-box optimization approaches like NES (no oracle DSWs required).
- Ideas for future work:

Conclusions and outlook

- A DSW-based audiovisual speaker tracking system can benefit from black-box optimization approaches like NES (no oracle DSWs required).
- Ideas for future work:
  - Making the system trainable end-to-end.

Conclusions and outlook

- A DSW-based audiovisual speaker tracking system can benefit from black-box optimization approaches like NES (no oracle DSWs required).
- Ideas for future work:
  - Making the system trainable end-to-end.
  - Joint optimization of DSW estimators and model parameters.

Conclusions and outlook

- A DSW-based audiovisual speaker tracking system can benefit from black-box optimization approaches like NES (no oracle DSWs required).
- Ideas for future work:
  - Making the system trainable end-to-end.
  - Joint optimization of DSW estimators and model parameters.
  - Extension to multi-speaker scenarios.

Conclusions and outlook

- A DSW-based audiovisual speaker tracking system can benefit from black-box optimization approaches like NES (no oracle DSWs required).
- Ideas for future work:
  - Making the system trainable end-to-end.
  - Joint optimization of DSW estimators and model parameters.
  - Extension to multi-speaker scenarios.

RUHR UNIVERSITÄT BOCHUM RUB

Thank you for your attention!