RUHR-UNIVERSITÄT BOCHUM

Cognitive
Signal Processing

RUB

Extending Linear Dynamical Systems with Dynamic Stream Weights for Audiovisual Speaker Localization

Christopher Schymura, Tobias Isenberg and Dorothea Kolossa

Cognitive Signal Processing Group, Institute of Communication Acoustics, Ruhr-University Bochum, Germany

Introduction

An important aspect of audiovisual speaker localization is the appropriate fusion of acoustic and visual observations based on their time-varying reliability. This study extends speaker localization and tracking frameworks based on linear dynamical systems with dynamic stream weights. A similar strategy has already been utilized for hidden Markov models in the context of audiovisual automatic speech recognition [1]. This work shows that dynamic stream weights can be integrated into the Kalman filter state estimation framework for linear dynamical systems and proposes a computationally efficient method for computing the corresponding acoustic and visual Kalman gains. Additionally, a method for estimating oracle dynamic stream weights from observation sequences with known state trajectories is introduced.

State Estimation

Linear dynamical system with conditionally independent acoustic and visual observations:

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{A}_k \boldsymbol{x}_{k-1} + \boldsymbol{v}_k, & \boldsymbol{v}_k &\sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{Q}_k) \\ \boldsymbol{y}_{\mathsf{A},k} &= \boldsymbol{C}_{\mathsf{A},k} \boldsymbol{x}_k + \boldsymbol{w}_{\mathsf{A},k}, & \boldsymbol{w}_{\mathsf{A},k} &\sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{R}_{\mathsf{A},k}) \\ \boldsymbol{y}_{\mathsf{V},k} &= \boldsymbol{C}_{\mathsf{V},k} \boldsymbol{x}_k + \boldsymbol{w}_{\mathsf{V},k}, & \boldsymbol{w}_{\mathsf{V},k} &\sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{R}_{\mathsf{V},k}) \end{aligned}$$

A recursive state estimation framework similar to the conventional Kalman filter can be derived for this class of linear dynamical systems, using the joint likelihood function at discrete time-step k with dynamic stream weights λ_k

$$\begin{aligned} &p(\boldsymbol{x}_0,\, \ldots,\, \boldsymbol{x}_k,\, \boldsymbol{y}_{\mathsf{A},1},\, \ldots,\, \boldsymbol{y}_{\mathsf{A},k},\, \boldsymbol{y}_{\mathsf{V},1},\, \ldots,\, \boldsymbol{y}_{\mathsf{V},k}) \propto \\ &\qquad p(\boldsymbol{x}_0) \prod_{k'=1}^{k} p(\boldsymbol{x}_{k'} \,|\, \boldsymbol{x}_{k'-1}) p(\boldsymbol{y}_{\mathsf{A},k'} \,|\, \boldsymbol{x}_{k'})^{\lambda_{k'}} p(\boldsymbol{y}_{\mathsf{V},k'} \,|\, \boldsymbol{x}_{k'})^{1-\lambda_{k'}} \end{aligned} \quad (1)$$

with

$$\begin{aligned} p(\boldsymbol{x}_0) &= \mathcal{N}(\boldsymbol{\mu}_0 \,|\, \boldsymbol{\Sigma}_0), & p(\boldsymbol{x}_k \,|\, \boldsymbol{x}_{k-1}) &= \mathcal{N}(\boldsymbol{A}_k \boldsymbol{x}_{k-1} \,|\, \boldsymbol{Q}_k) \\ p(\boldsymbol{y}_{\mathsf{A},k} \,|\, \boldsymbol{x}_k) &= \mathcal{N}(\boldsymbol{C}_{\mathsf{A},k} \boldsymbol{x}_k \,|\, \boldsymbol{R}_{\mathsf{A},k}), & p(\boldsymbol{y}_{\mathsf{V},k} \,|\, \boldsymbol{x}_k) &= \mathcal{N}(\boldsymbol{C}_{\mathsf{V},k} \boldsymbol{x}_k \,|\, \boldsymbol{R}_{\mathsf{V},k}) \end{aligned}$$

Require: Observations $\{\boldsymbol{y}_{\mathsf{A},k}\}_{k=1}^{K}$, $\{\boldsymbol{y}_{\mathsf{V},k}\}_{k=1}^{K}$, dynamic stream weights $\{\lambda_k\}_{k=1}^{K}$, initial state mean and covariance matrix $\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{\Sigma}}_0$.

for $k = 1$ to K **do**

// Prediction step

$\hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{A}_k \hat{\boldsymbol{x}}_{k-1}$

$\hat{\boldsymbol{\Sigma}}_{k|k-1} = \boldsymbol{A}_k \hat{\boldsymbol{\Sigma}}_{k-1} \boldsymbol{A}_k^{\mathsf{T}} + \boldsymbol{Q}_k$

// Compute acoustic and visual Kalman gain

$\boldsymbol{S}_{\mathsf{A},k} = \boldsymbol{R}_{\mathsf{A},k}^{-1} \boldsymbol{C}_{\mathsf{A},k} \hat{\boldsymbol{\Sigma}}_{k|k-1}$; $\boldsymbol{S}_{\mathsf{V},k} = \boldsymbol{R}_{\mathsf{V},k}^{-1} \boldsymbol{C}_{\mathsf{V},k} \hat{\boldsymbol{\Sigma}}_{k|k-1}$

$$\begin{bmatrix} \boldsymbol{K}_{\mathsf{A},k}^{\mathsf{T}} \\ \boldsymbol{K}_{\mathsf{V},k}^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} + \lambda_k \boldsymbol{S}_{\mathsf{A},k} \boldsymbol{C}_{\mathsf{A},k}^{\mathsf{T}} & (1-\lambda_k) \boldsymbol{S}_{\mathsf{A},k} \boldsymbol{C}_{\mathsf{V},k}^{\mathsf{T}} \\ \lambda_k \boldsymbol{S}_{\mathsf{V},k} \boldsymbol{C}_{\mathsf{A},k}^{\mathsf{T}} & \boldsymbol{I} + (1-\lambda_k) \boldsymbol{S}_{\mathsf{V},k} \boldsymbol{C}_{\mathsf{V},k}^{\mathsf{T}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{S}_{\mathsf{A},k} \\ \boldsymbol{S}_{\mathsf{V},k} \end{bmatrix}$$

// Update step

$\tilde{\boldsymbol{y}}_{\mathsf{A},k} = \boldsymbol{y}_{\mathsf{A},k} - \boldsymbol{C}_{\mathsf{A},k} \hat{\boldsymbol{x}}_{k|k-1}$; $\tilde{\boldsymbol{y}}_{\mathsf{V},k} = \boldsymbol{y}_{\mathsf{V},k} - \boldsymbol{C}_{\mathsf{V},k} \hat{\boldsymbol{x}}_{k|k-1}$

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

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

end for

return estimated state trajectory $\{\hat{\boldsymbol{x}}_1,\, \ldots,\, \hat{\boldsymbol{x}}_K\}$

Algorithm 1: State estimation algorithm based on the conventional Kalman filter recursions, which are extended to cope with dynamic stream weights in linear dynamical systems. The mathematical derivation is based on the MMSE criterion using the joint likelihood function with dynamic stream weights in Eq. (1).

Oracle Dynamic Stream Weights

If a dataset with audiovisual observation sequences $\boldsymbol{Y}_{\mathsf{A}} = \{\boldsymbol{y}_{\mathsf{A},k}\}_{k=1}^{K}$, $\boldsymbol{Y}_{\mathsf{V}} = \{\boldsymbol{y}_{\mathsf{V},k}\}_{k=1}^{K}$ and the corresponding true state sequence of speaker positions $\boldsymbol{X} = \{\boldsymbol{x}_k\}_{k=1}^{K}$ is available, the sequence of oracle dynamic stream weights $\boldsymbol{\lambda}^{\star} = \{\lambda_k\}_{k=1}^{K}$ can be computed according to

$$\boldsymbol{\lambda}^{\star} = \arg \max_{\lambda_1,\, \ldots,\, \lambda_K} \prod_{k=1}^{K} p(\lambda_k)\, p(\boldsymbol{y}_{\mathsf{A},k} \,|\, \boldsymbol{x}_k)^{\lambda_k}\, p(\boldsymbol{y}_{\mathsf{V},k} \,|\, \boldsymbol{x}_k)^{1-\lambda_k}.$$

By assuming i.i.d. oracle stream weights and imposing a Gaussian prior with mean μ_λ and variance σ_λ^2, a closed-form solution can be obtained:

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

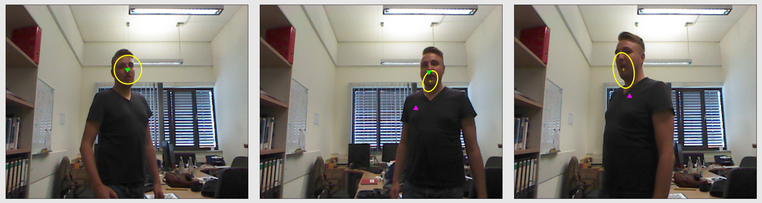

Figure 1: Exemplary speaker localization and tracking situations using oracle dynamic stream weights. Video observations are shown as green and audio observations as magenta triangles. The estimated state mean and covariance matrix are depecited in yellow.

Evaluation

Speaker localization experiments were conducted using 150 audiovisual recordings of the humanoid robot Nao in a reverberant room with $T_{60} \approx 450\,\text{ms}$. DPD-MUSIC [2] was used for acoustic localization and visual speaker locations were estimated with the Viola-Jones algorithm [3]. The evaluation metric is localization RMSE. Systematic errors in visual localization were generated by adding zero-mean Gaussian noise at different standard deviations.

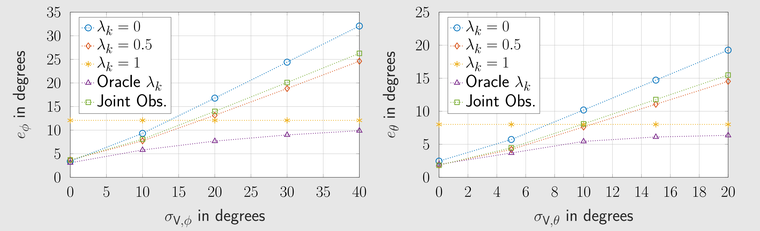

Figure 2: Localization errors obtained with different dynamic stream weight settings at varying levels of disturbance for the visual modality. Azimuth errors are shown in the left plot, elevation errors are depicted in the right plot. "Joint Obs." shows the baseline performance obtained with a conventional Kalman filter.

References

[1] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "A new EM estimation of dynamic stream weights for coupled-HMM-based audio-visual ASR," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[2] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.