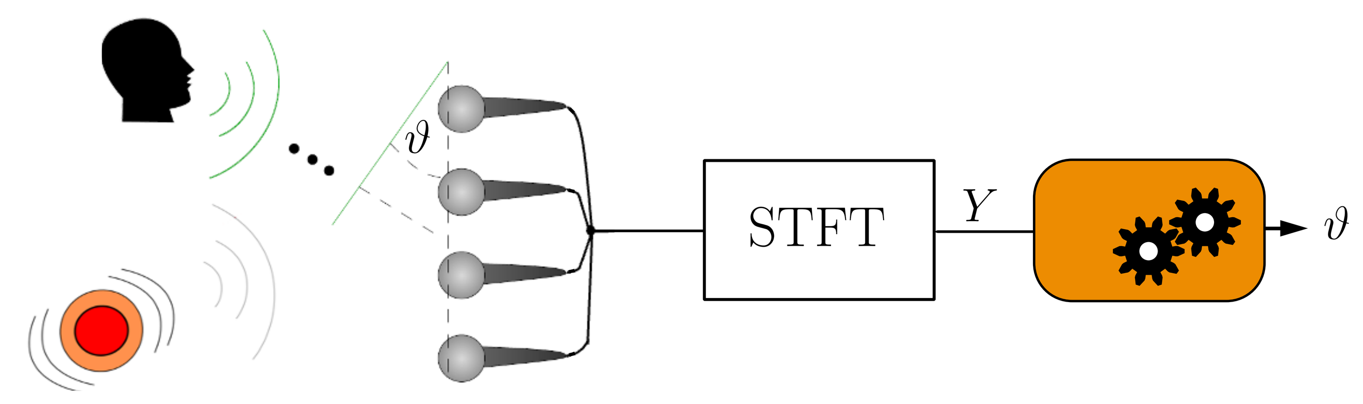
End-to-End Signal-Aware Direction-of-Arrival Estimation Using Weighted Steered-Response Power

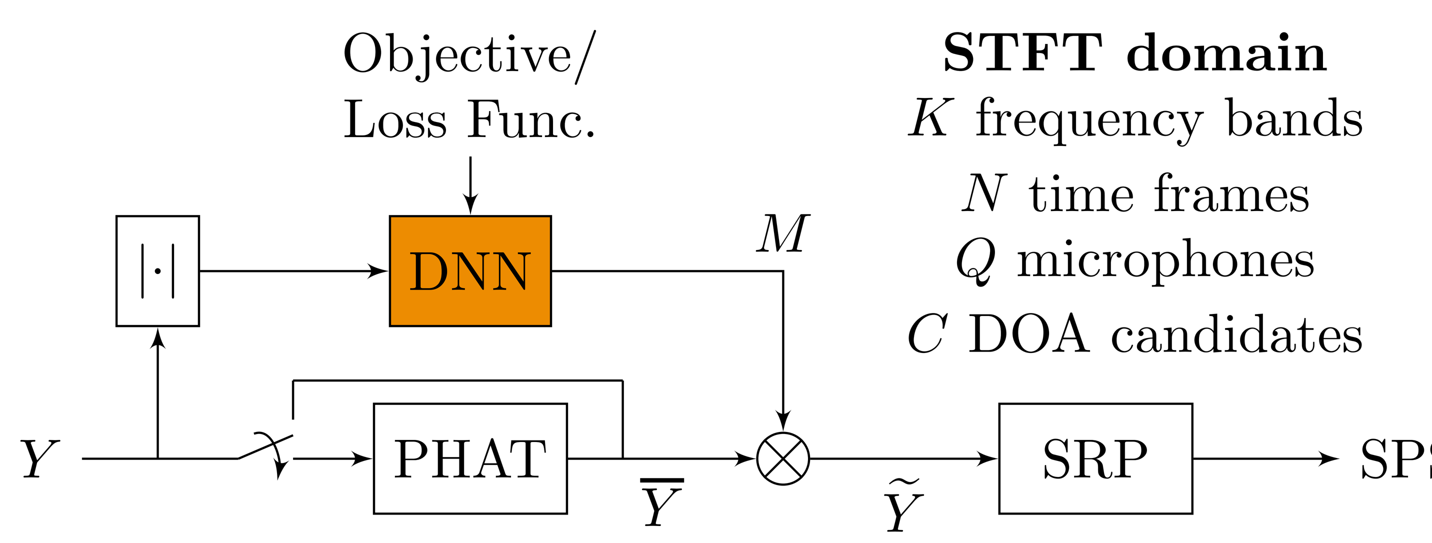
Julian Wechsler, Wolfgang Mack, Emanuël A. P. Habets

1. Introduction – Signal-Aware DOA Estimation



- Estimate the **direction-of-arrival (DOA)** ϑ **of a desired source type** using a uniform linear microphone array with Q microphones
- Enable, e.g., automatic steering of a camera towards a speaker

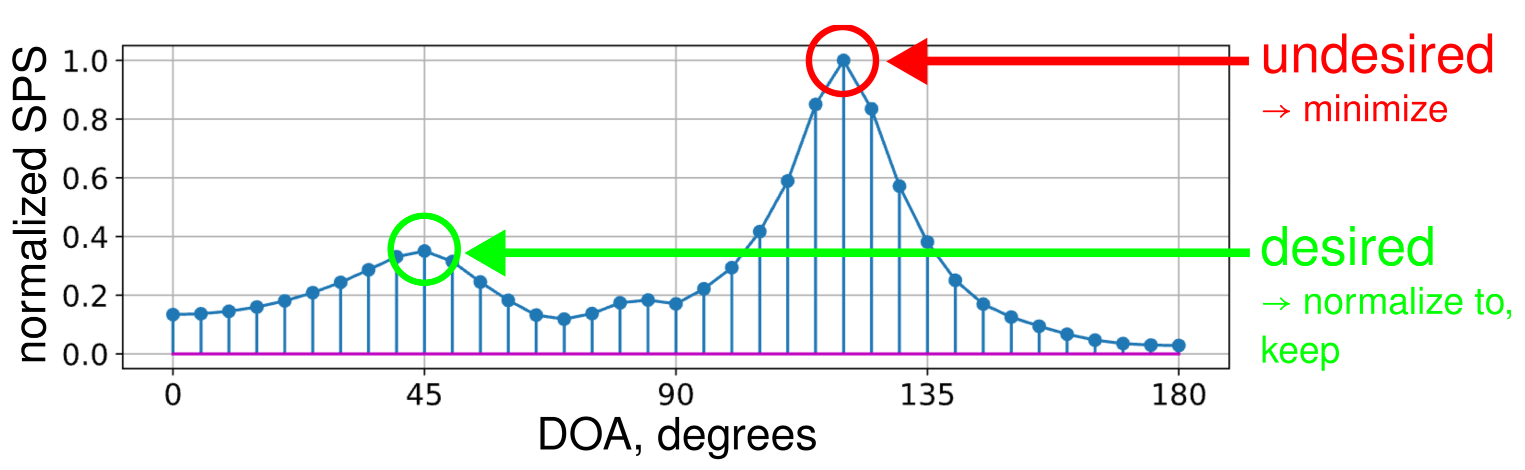
2. Problem Formulation



$Y, \overline{Y}, \widetilde{Y} \in \mathbb{C}^{K\times N\times Q}$ $\quad M \in [0,1]^{K\times N\times Q}$ $\quad \text{SPS} \in \mathbb{R}^{C\times N}$

- Microphone signals Y: Mixture of direct components, reverberation, and microphone self noise; desired **DOA information encapsulated in phase component of the direct sound of the desired source**
- DOA estimation using **weighted Steered-Response Power (SRP)**:
  - **Selection of time-frequency bins supporting desired DOA estimate** by deep neural network (DNN)-based mask M
  - **General robustification** by phase transform (PHAT) weighting [1]
- **Hybrid approach**: Based on DNNs and classical signal processing
  - Existing loss functions for hybrid systems **require the direct sound of the source of interest** as training reference
  - E.g., using an mean squared error loss, the phase-sensitive mask (PSM) was shown to improve signal-aware DOA estimation [2]

3. Proposed Training Strategy



- We propose a **solely DOA-based end-to-end loss** for hybrid systems, based on the spatial pseudo-spectrum (SPS)
- Idea: **Minimize the overall output power while retaining it from the direction of the source of interest (SOI)**
- The **power minimization loss (PML)** is based on the SPS obtained from the microphone signals after PHAT weighting and masking, $\widetilde{Y}$
- For time frame $n \in \{1, 2, \ldots, N\}$ and DOA candidate $c \in \{1, 2, \ldots, C\}$, and with the knowledge of the DOA of the SOI, c_{SOI}, we define

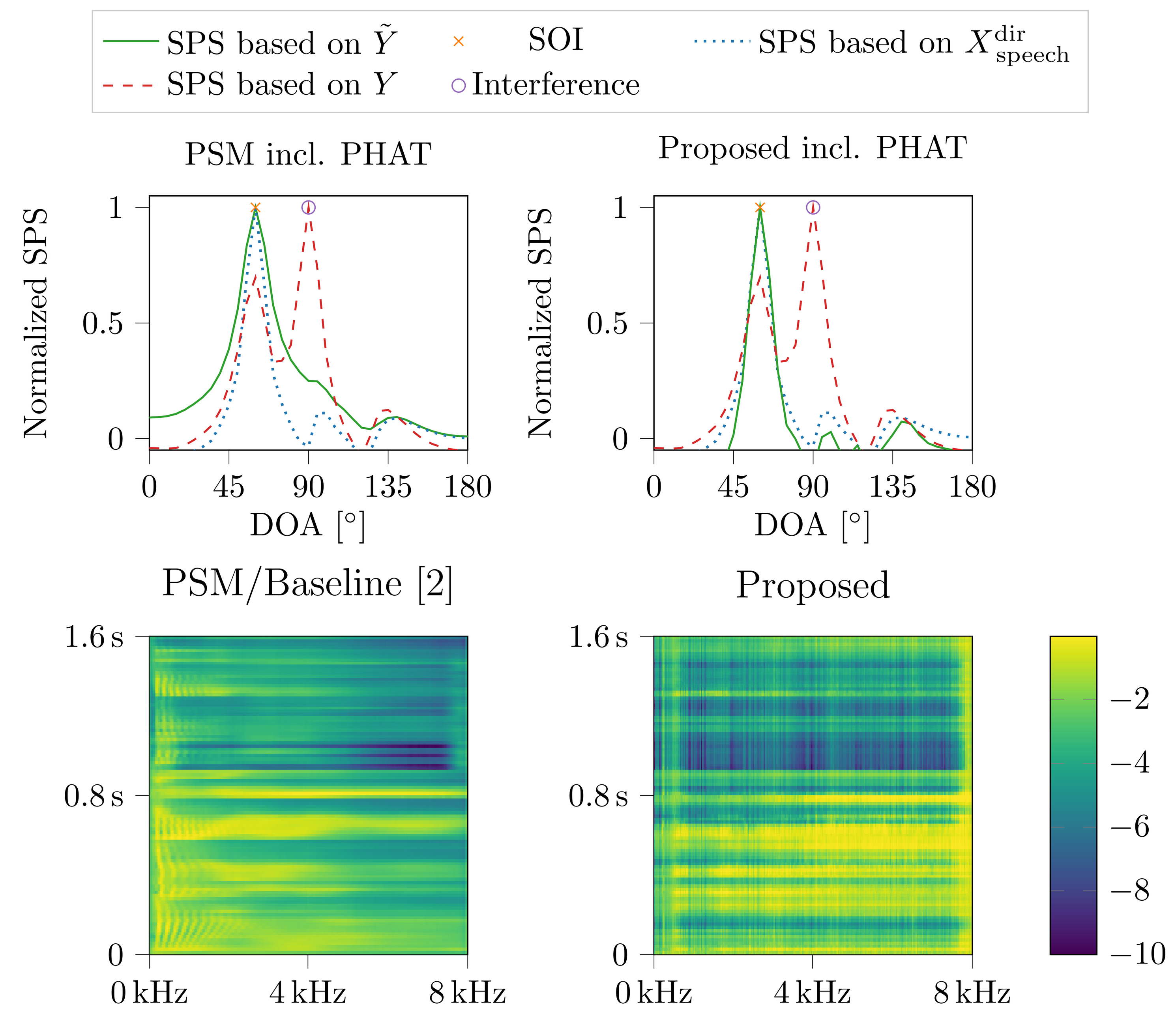
$$\text{PML} = \frac{1}{\sum_{n=1}^{N} \text{SPS}\left(\widetilde{Y}\right)[c_{\text{SOI}}, n]} \sum_{c=1}^{C}\sum_{n=1}^{N} \text{SPS}\left(\widetilde{Y}\right)[c, n]$$

4. Experimental Setup (from [3])

- **Trained with simulated room impulse responses** of 5 rooms with 5 different reverberation times each, random positions for the array
- **Tested with measured room impulse responses**
- **STFT parameters**: sampling frequency 16 kHz, window length 32 ms, hop size 16 ms, Hann window
- $K = 257$, $N = 100$, $Q = 4$, $C = 37$
- **DNN**: 2 LSTM layers (hidden dim. = 512), 1 feed-forward layer with sigmoid activation

5. Performance Evaluation

- **Proposed method accomplished state-of-the-art performance** reducing the mean absolute error from ~40° to 7.8° (baseline: 6.8°)
- Time-frequency structure of speech not visible in end-to-end mask



6. Conclusions

- Training of hybrid signal-aware DOA estimation system w/o access to direct sound of sources, enabling training on measured data
- **Proposed method achieves state-of-the-art performance**
- End-to-end mask can focus on high SIR regions, but is not usable for speech enhancement purposes

[1] Joseph Hector DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University Providence, RI, May 2000.

[2] Zhong Wang, Xueliang Zhang, and DeLiang Wang, “Robust speaker localization guided by deep learning-based time-frequency masking,” *IEEE/ACM Trans. Aud., Sp., Lang. Proc.*, vol. 27, no. 1, pp. 178–188, 2019.

[3] Wolfgang Mack, Julian Wechsler, and Emanuël A. P. Habets, “End-to-end signal-aware direction-of-arrival estimation using attention mechanisms,” *Computer Speech & Language*, vol. 75, pp. 101363, 2022.

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg
Fraunhofer IIS
ASC
Elitenetzwerk Bayern