FACULTY OF ENGINEERING AND ARCHITECTURE

Exploiting Temporal Context in CNN Based Multisource DOA Estimation

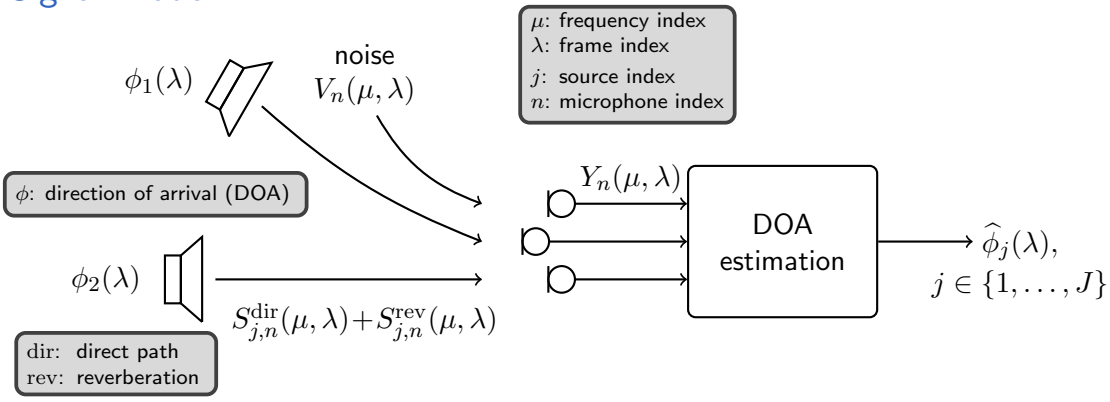
Alexander Bohlender[1], Ann Spriet[2], Wouter Tirry[2], and Nilesh Madhu[1]

[1] IDLab, Department of Electronics and Information Systems, Ghent University - imec, Ghent, Belgium
[2] Goodix Technology (Belgium) B.V., Leuven, Belgium

International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2022

GHENT UNIVERSITY imec

Signal Model



$$Y_n(\mu, \lambda) = \sum_j \Big(S^{\text{dir}}_{j,n}(\mu, \lambda) + S^{\text{rev}}_{j,n}(\mu, \lambda) \Big) \; + \; V_n(\mu, \lambda)$$

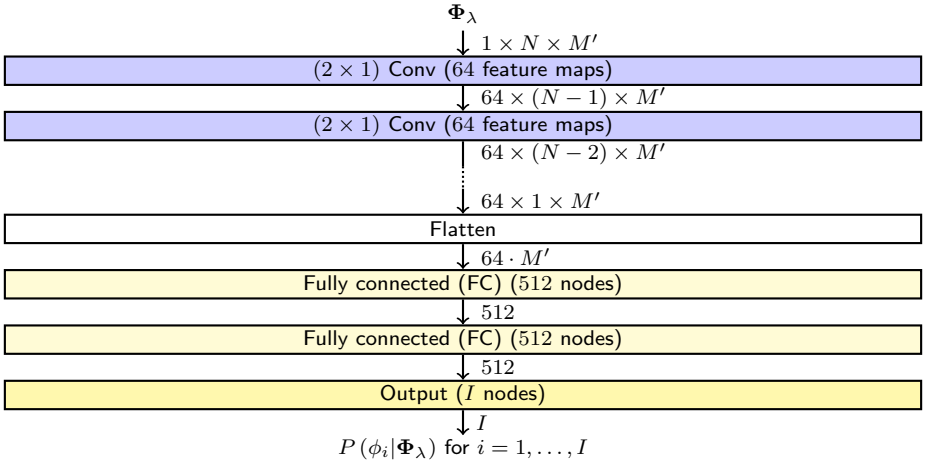
The DOA ϕ can be an angle (azimuth φ) or a set of angles (azimuth φ and elevation ϑ).

CNN Based (Single-Frame) DOA Classification[1]

- The (unavailable) direct-path components $S_{j,n}^{\mathrm{dir}}(\mu, \lambda)$ are attenuated and phase-shifted versions of the clean source signals, where the phase shifts are dependent on the DOA.
  - Therefore, the (available) **microphone signal phases** $\angle Y_n(\mu, \lambda)$ represent a suitable **input** for a deep neural network (DNN).
  - Define a *phase map* $\boldsymbol{\Phi}_\lambda$, which consists of the phases for the M' discrete frequencies up to the Nyquist frequency for all N microphones.
- We can interpret DOA estimation as a classification problem, where the classes are defined based on a grid of I discrete DOAs $\phi \in \{\phi_1, \ldots, \phi_I\}$.
  - The **posterior probabilities** $P(\phi_i | \boldsymbol{\Phi}_\lambda)$ represent a suitable **output** for a DNN.

[1] S. Chakrabarty and E. A. P. Habets. "Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained With Noise Signals". In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (2019), pp. 8–21. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2019.2901664

CNN Based (Single-Frame) DOA Classification[1]: Architecture



[1] S. Chakrabarty and E. A. P. Habets. "Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained With Noise Signals". In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (2019), pp. 8–21. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2019.2901664

CNN Based (Single-Frame) DOA Classification[1]: Training Data

$$Y_n(\mu,\lambda) = \sum_j \Big(S^{\mathrm{dir}}_{j,n}(\mu,\lambda) + S^{\mathrm{rev}}_{j,n}(\mu,\lambda) \Big) + V_n(\mu,\lambda) = \sum_j S^{\mathrm{mic}}_{j,n}(\mu,\lambda) + V_n(\mu,\lambda)$$

Sources $S^{\mathrm{mic}}_{j,n}(\mu,\lambda)$:	time domain convolution of uncorrelated noise with simulated room room impulse responses
DOAs:	a fixed direction is randomly selected for each source
Summation:	synthetically enforced W-disjoint orthogonality (only 1 source per time-frequency bin contributes to the microphone signals) ➔ DNN learns to make use of the approximate W-disjoint orthogonality of speech
Noise $V_n(\mu,\lambda)$:	noise (spatially and temporally uncorrelated) is added with a randomly chosen signal-to-noise ratio

[1] S. Chakrabarty and E. A. P. Habets. "Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained With Noise Signals". In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (2019), pp. 8–21. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2019.2901664

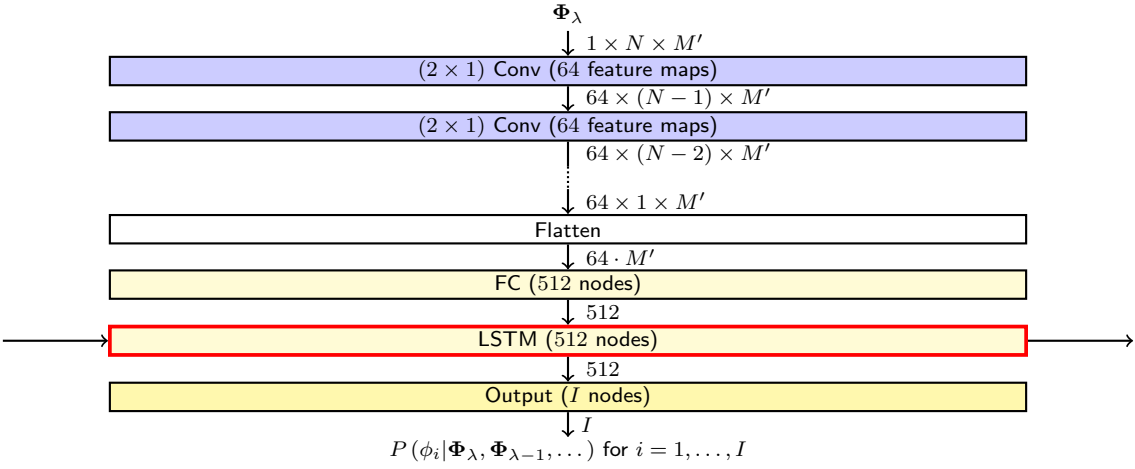
Extension to Multi-Frame DOA Estimation

1. The DOA estimation is performed independently for each frame, although temporal context is very useful particularly due to the DOAs typically changing quite slowly.
   - We can **modify the architecture** so that information from previous frames can be taken into account as well.
2. The training data generation is designed for single-frame DOA estimation.
   - For an extension to multi-frame DOA estimation, we propose a model based **generation of simulated training data** that accounts for time-variant source activity and DOA changes.
3. To make the best use of the approach, a better understanding of the relation between training setup and performance is needed.
   - In addition to comparing different variants of the architecture, we also conduct experiments to **evaluate the importance of various parameters**, including the source signal type and the spatial characteristics of the noise.

The aim of this work is to study how best to incorporate temporal context in DNN based DOA estimation *in general*, and how training data suitable for this purpose can be generated accordingly, not the specific CNN approach we make use of.

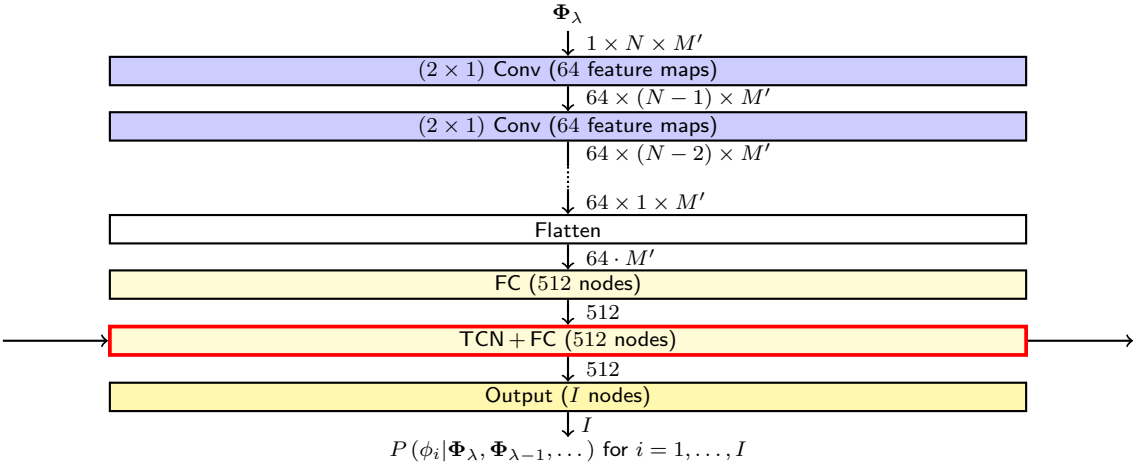
CNN Architectures for DOA Classification With Temporal Context

Replace the second FC layer by a long short-term memory (LSTM) layer:



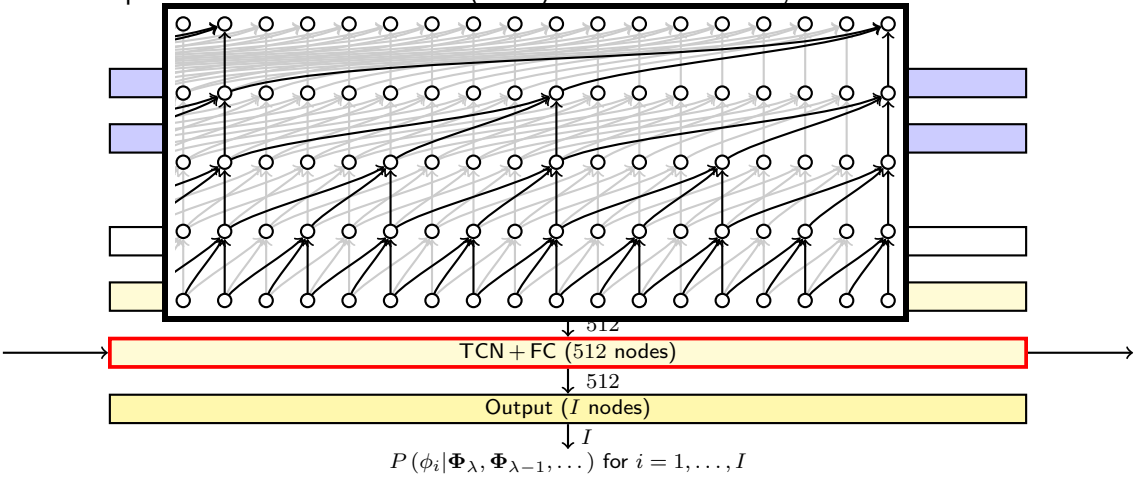
CNN Architectures for DOA Classification With Temporal Context

Add a temporal convolutional network (TCN) between the FC layers:



CNN Architectures for DOA Classification With Temporal Context

Add a temporal convolutional network (TCN) between the FC layers:

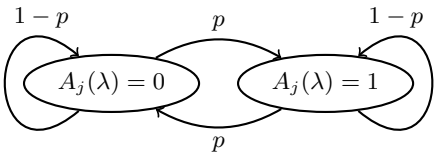


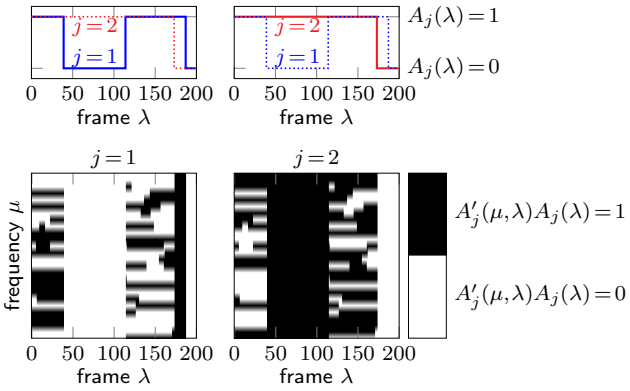
Training Data: Simulating Dynamic Scenes

Updated signal model:

only needed when noise is used for the source signals!

$$Y_n(\mu,\lambda) = \sum_j \left(A'_j(\mu,\lambda)\, A_j(\lambda)\, S^{\text{mic}}_{j,n}(\mu,\lambda)\right) + V_n(\mu,\lambda)$$





Training Data: Generating the Microphone Signals

$$Y_n(\mu, \lambda) = \sum_j \left(A_j(\lambda)\, S_{j,n}^{\text{mic}}(\mu, \lambda) \right) + V_n(\mu, \lambda)$$

Sources $S_{j,n}^{\text{mic}}(\mu, \lambda)$: time domain convolution of clean speech with simulated room room impulse responses

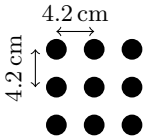
DOAs: a different direction is randomly selected every time a source becomes active again $(A_j(\lambda) = 1, A_j(\lambda - 1) = 0)$

Summation: no need to artificially introduce W-disjointness when realistic source signals are used → sufficient to simply sum up all sources

Noise $V_n(\mu, \lambda)$: noise (spatially diffuse, temporally uncorrelated) is added with a randomly chosen signal-to-noise ratio

Evaluation Setup

$$Y_n(\mu, \lambda) = \sum_j S_{j,n}^{\text{mic}}(\mu, \lambda) + V_n(\mu, \lambda)$$

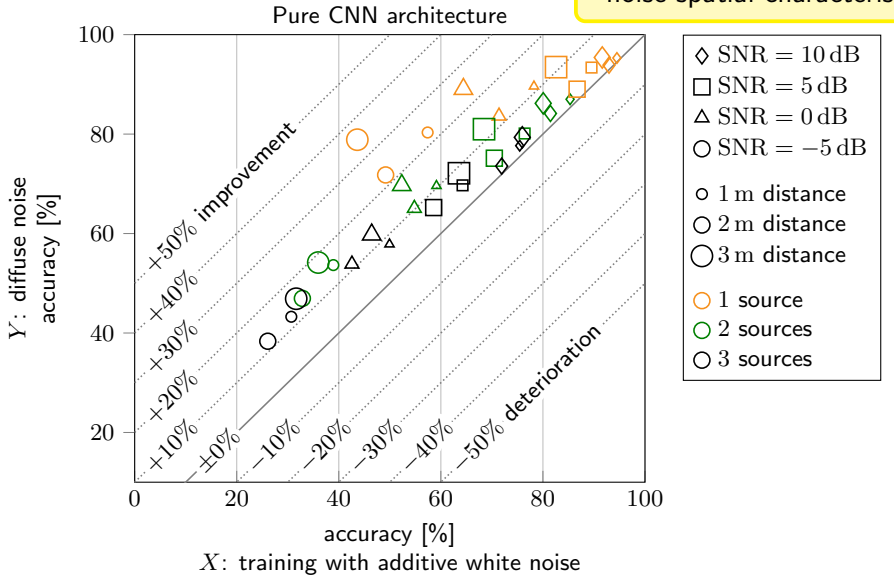


Sources $S_{j,n}^{\text{mic}}(\mu, \lambda)$:	time domain convolution of clean speech with recorded room room impulse responses ($T_{60} = 660\,\text{ms}$)
DOAs:	with a probability of 50%, a different direction is randomly selected after the end of an utterance
Noise $V_n(\mu, \lambda)$:	spatially diffuse, temporally uncorrelated $\bigoplus$
Microphone array:	9-microphone uniform rectangular array $\bigoplus$
DOA estimation:	$I = 37$ DOA classes (azimuth angles $\varphi = 0°, 5°, \ldots, 180°$)

$\bigoplus$ Additional results for a different setup available in the paper.

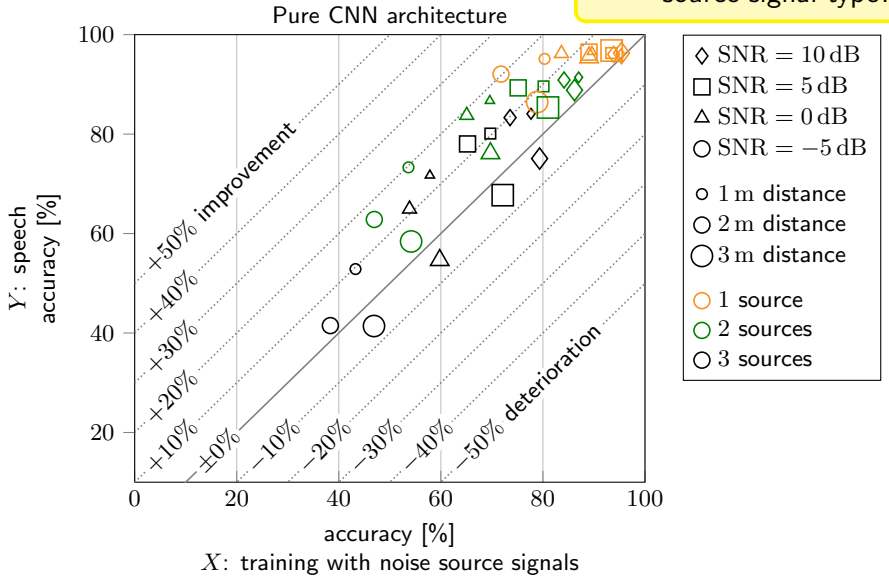
Experimental Results: Training Setup

noise spatial characteristics?



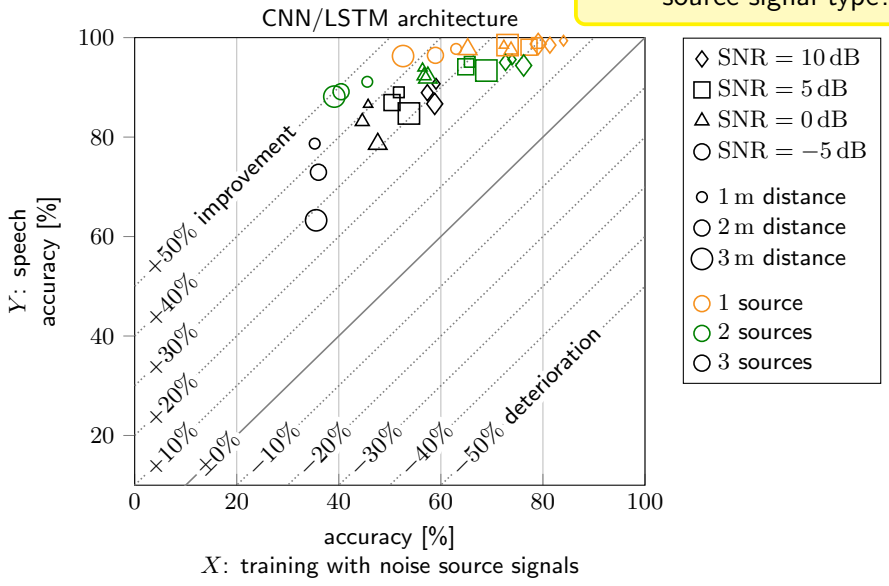
Experimental Results: Training Setup

source signal type?

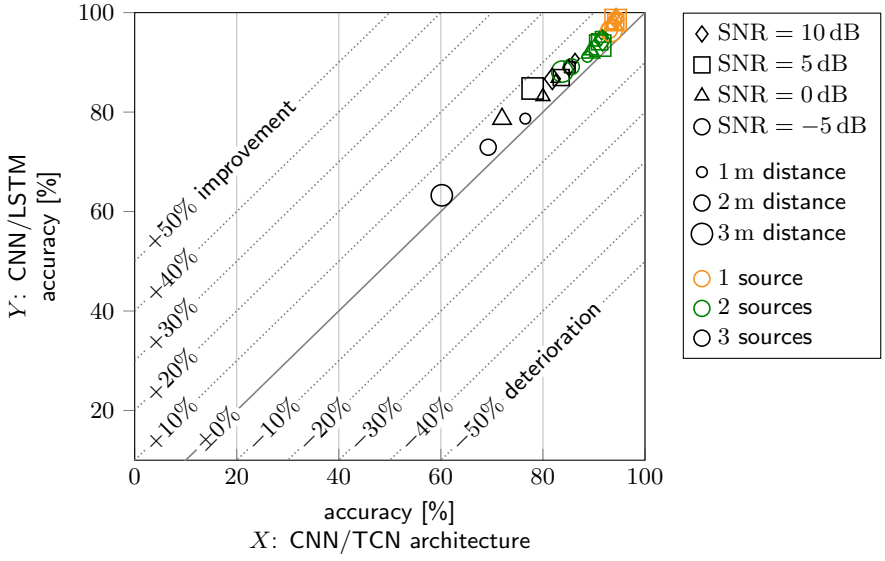


Experimental Results: Training Setup

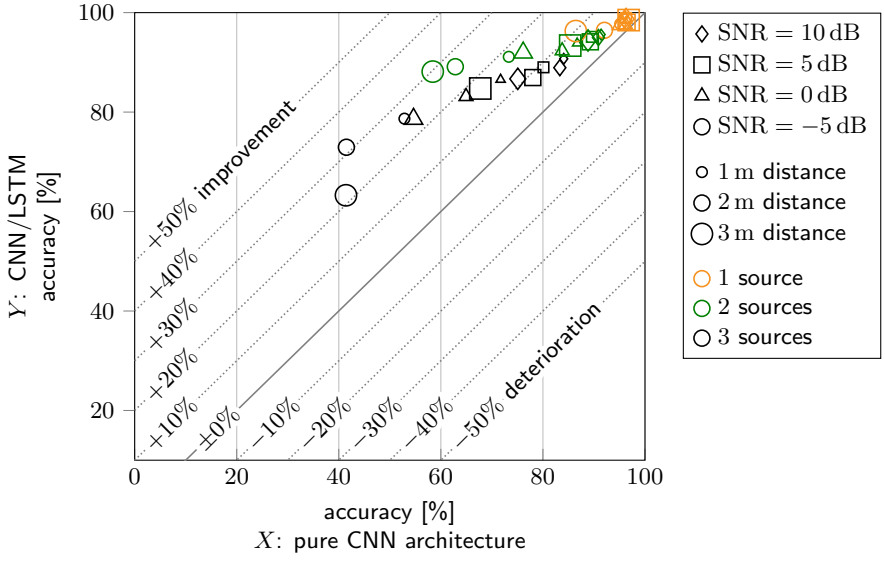
source signal type?



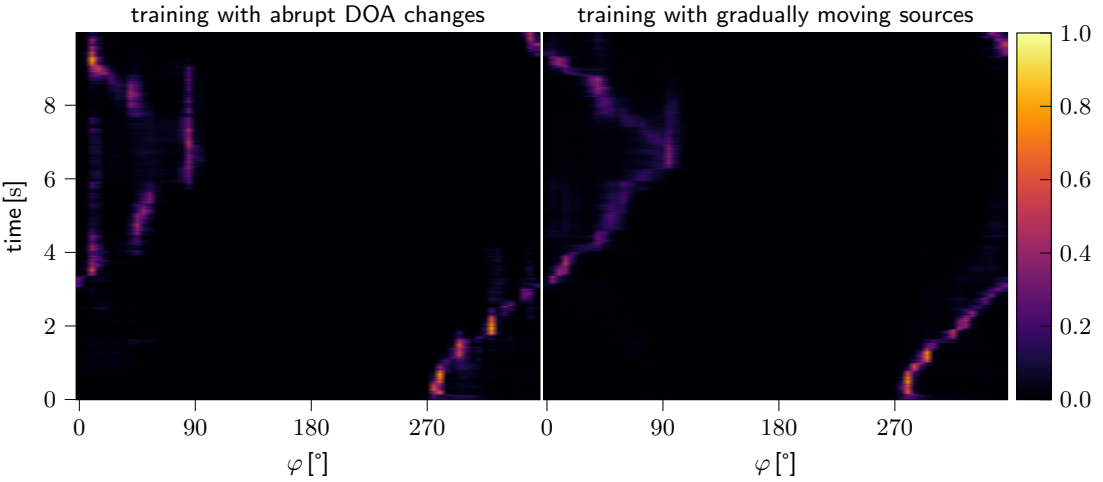
Experimental Results: Training: Architecture Extensions



Experimental Results: Training: Architecture Extensions



Moving Sources[1]



[1] Alexander Bohlender, Liesbeth Roelens, and Nilesh Madhu. *Latency Controlled Deep Moving Speaker Tracking Using Simulated Training Data*. unpublished (submission under review)

Conclusions

The following conclusions can be drawn with regard to the formulated research goals:

1. *The employed architecture must be capable of exploiting temporal dependencies.*
   - Relatively slowly changing DOAs make it interesting to take advantage of long-term temporal context.
   - A simple but effective option to achieve this is to include an LSTM layer in the CNN.
2. *The training data must allow the DNN to learn how the DOA estimation can be improved with information from previous frames. This includes the detection of DOA changes.*
   - We simulate acoustic scenes with time-variant source activity and DOAs.
   - This enables the CNN/LSTM approach to clearly outperform the CNN baseline, where the output is simply averaged over a fixed period of time.
3. *A better understanding of the relation between training setup and performance is needed.*
   - Experiments demonstrate that realistic (correlated) source signals should be used to train the network when the architecture permits the use of temporal context.
   - The spatial properties of the noise are relevant as well.