INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

AUDIO LABS

Broadband DOA Estimation using Convolutional Neural Networks Trained with Noise Signals

Soumitro Chakrabarty and Emanuël Habets

WASPAA 2017

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

Motivation
Signal processing methods

- Cross-correlation-based methods
  - GCC-PHAT
  - SRP-PHAT
  - MCCC…
- Subspace-based methods
  - MUSIC...
- Model-based methods
  - Maximum-likelihood estimation...
- ...

Challenges

- Performance degradation in presence of noise and reverberation
- High computational cost

Motivation
Supervised learning methods

- Advantage: Supervised learning methods can be adapted to different acoustic environments

- Recently, deep neural network (DNN) based supervised learning methods have been successful across a range of applications:
  - Automatic speech recognition
  - Object recognition in images
  - Machine translation….

- Few DNN based methods that estimate DOA of a sound source from the observed signals

[1] Xiao et al. "A learning-based approach to direction of arrival estimation in noisy and reverberant environments,“ 2015
[2] Takeda et al. "Sound source localization based on deep neural networks with directional activate function exploiting phase information“ 2016

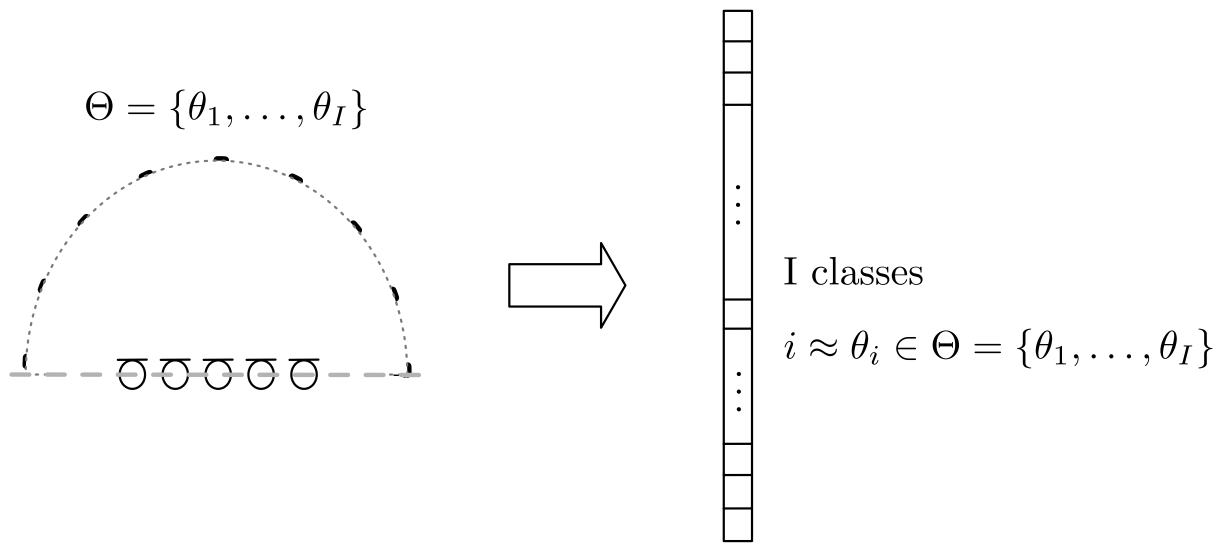
Motivation
DNN based DOA estimation

Aim: A DNN based supervised learning method for DOA estimation that

- Estimates DOA per time frame given the STFT representation of the observed signals
- Simple input representation to learn relevant features during training

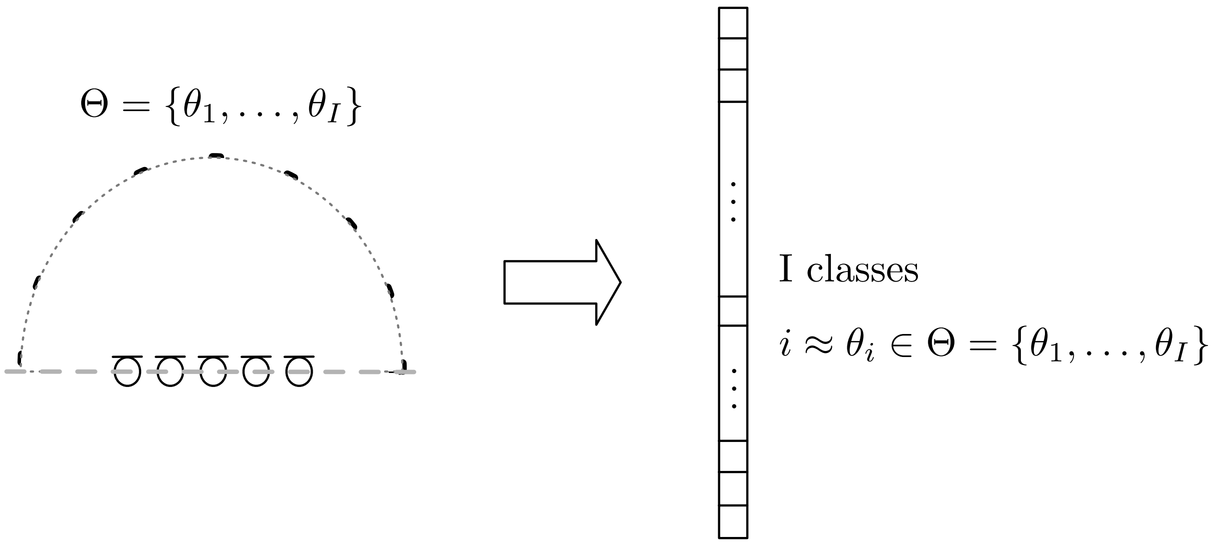
Problem Formulation

DOA estimation as classification



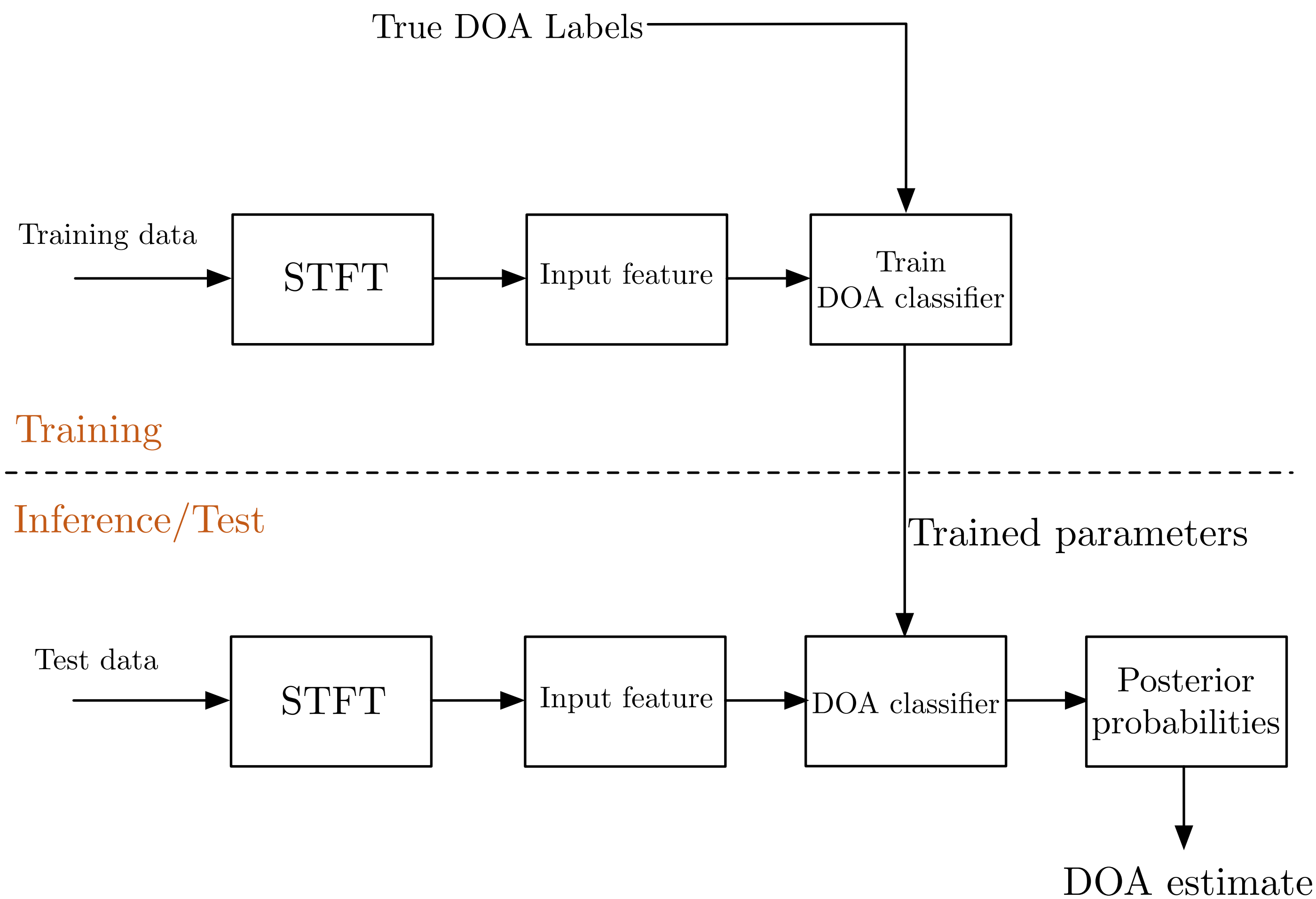
- DOA estimation is formulated as an I class classification problem
- Discretize the whole DOA range into I discrete values to obtain a set of possible DOA values: $\Theta = \{\theta_1, \ldots, \theta_I\}$
- Each class corresponds to a possible DOA value in the set

Problem Formulation
DOA estimation as classification

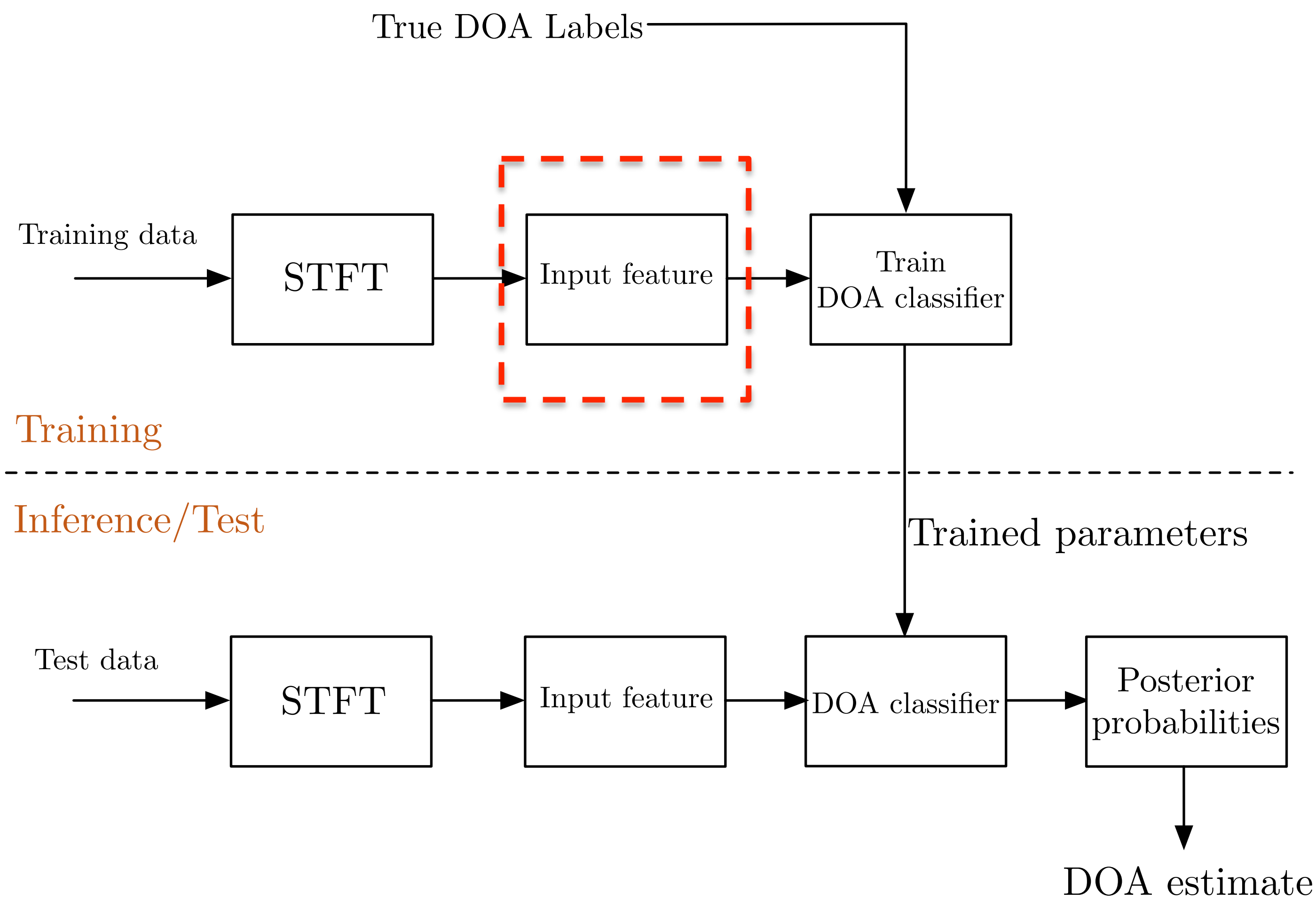


- For each frame, compute the posterior probability for each class
- DOA estimate is the DOA of the class with the highest posterior

System Overview
Supervised learning framework



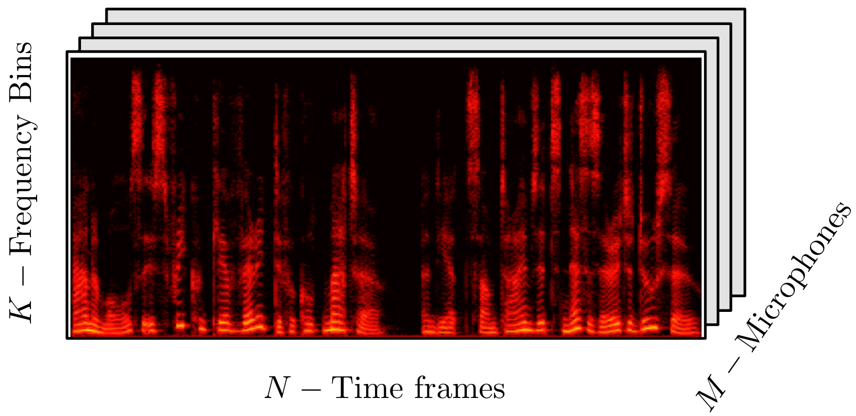
System Overview
Supervised learning framework



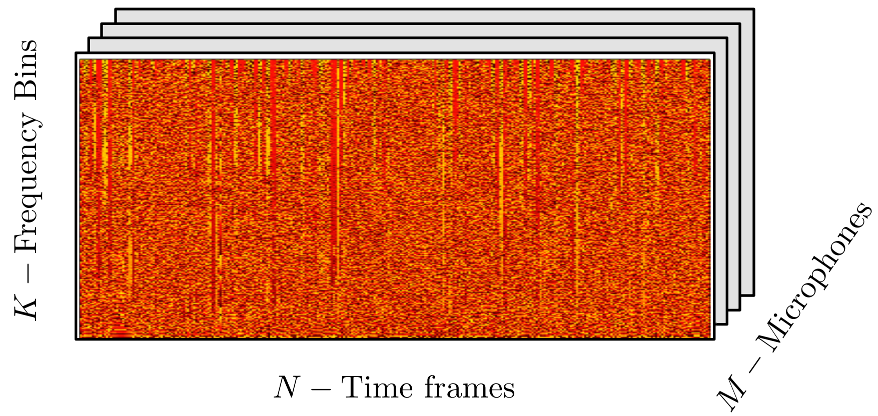
Input feature representation

STFT magnitude and phase component

$$Y_m(n,k) = A_m(n,k)e^{j\phi_m(n,k)}$$



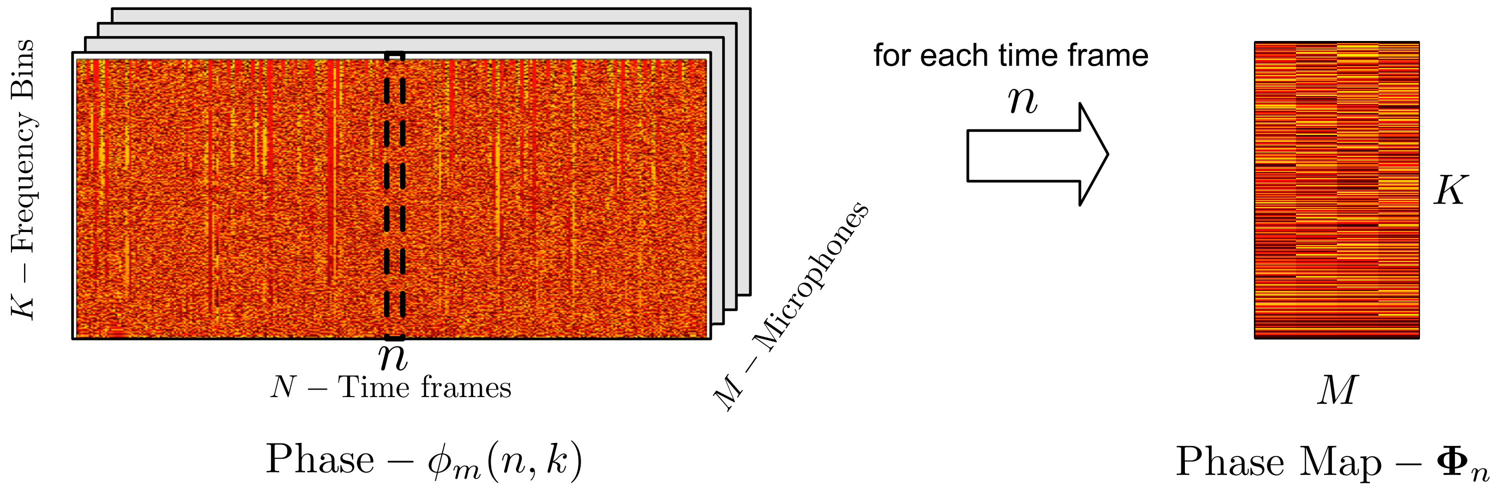
Magnitude – $A_m(n,k)$



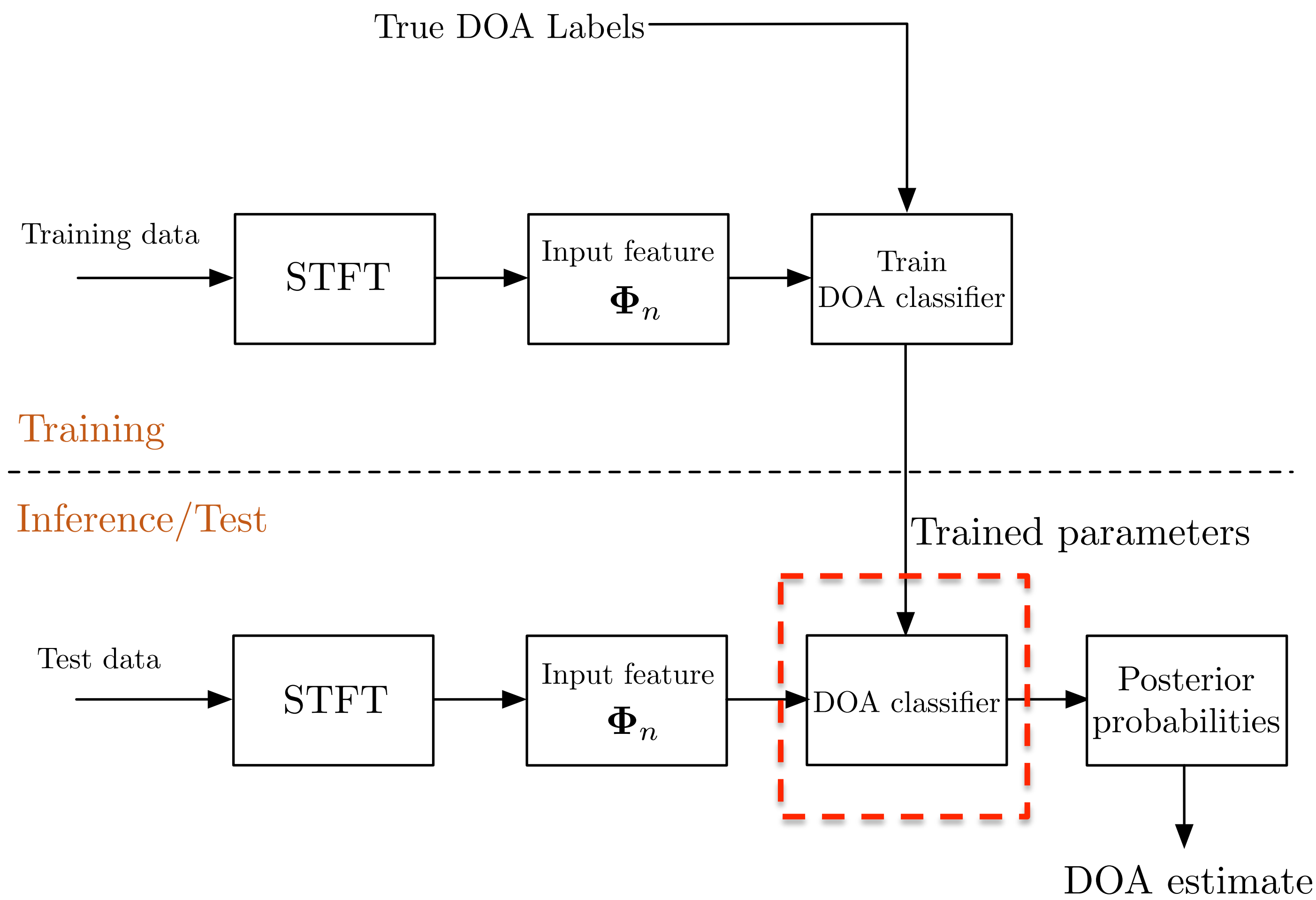
Phase – $\phi_m(n,k)$

Input feature representation

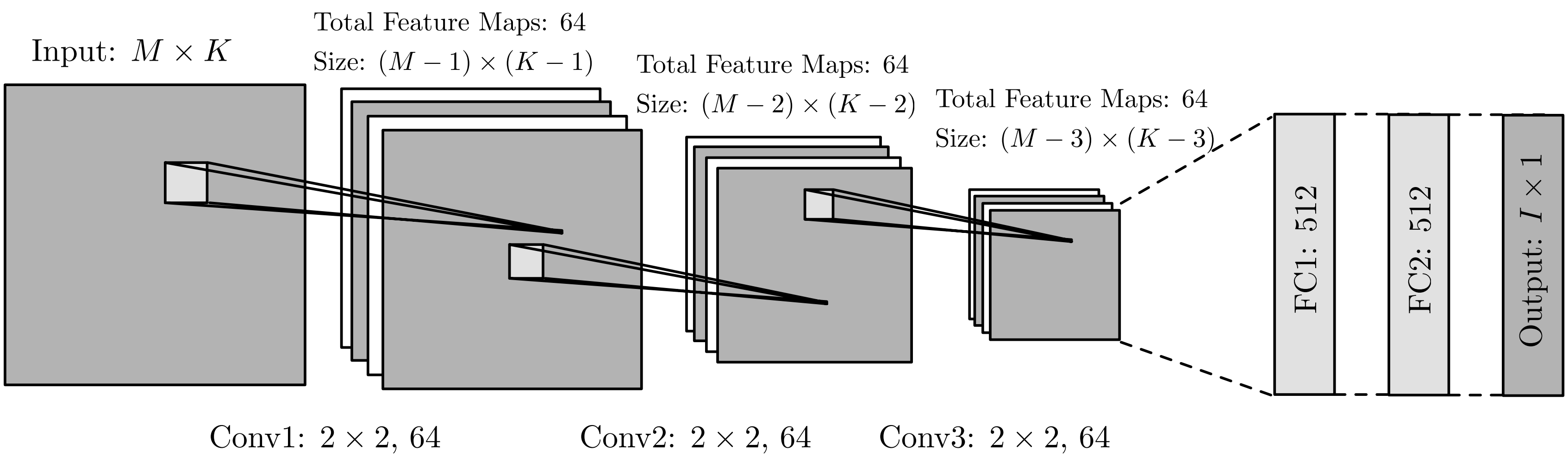
Phase map



System Overview
Supervised learning framework



CNN Architecture

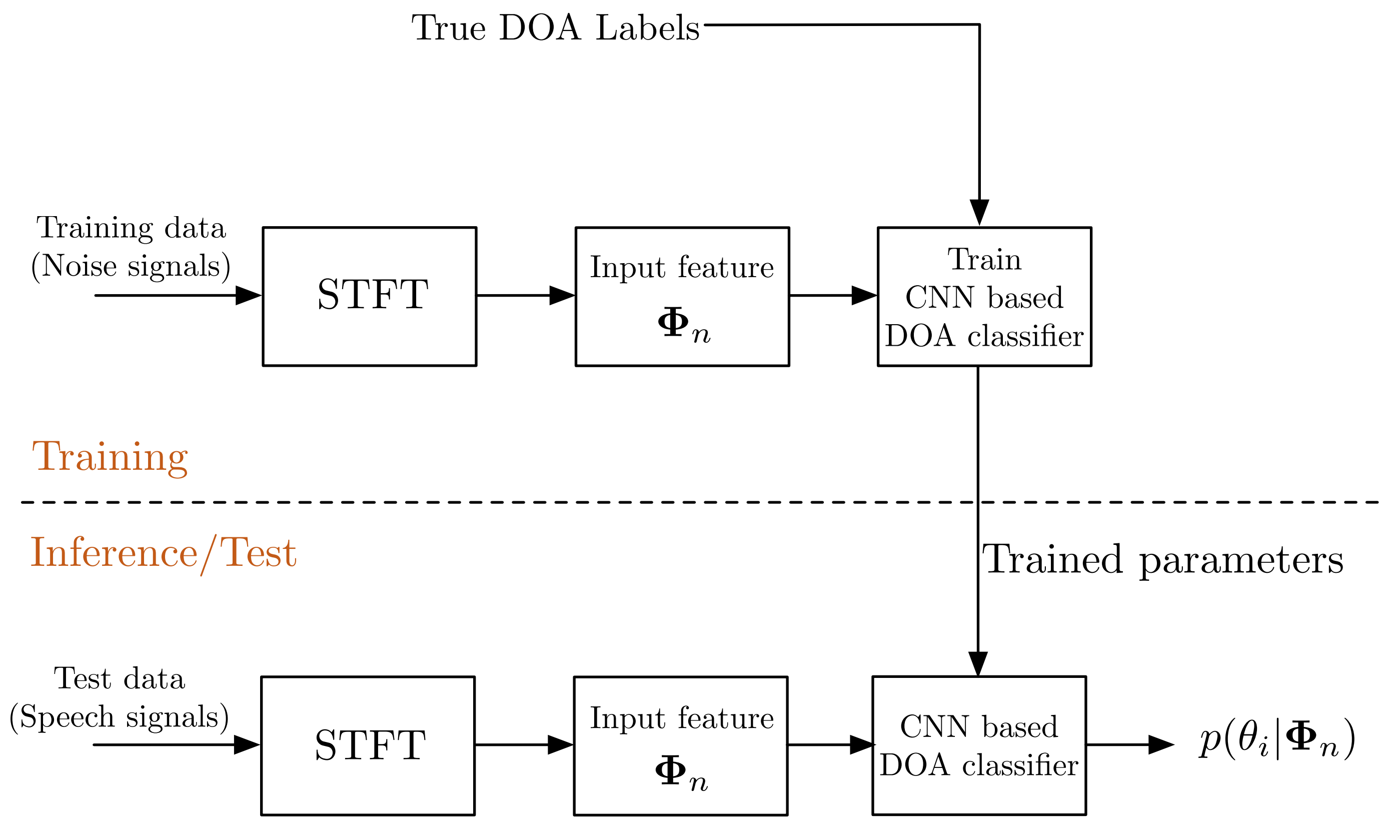


- Pooling is not employed
- Experiments showed decreased performance

Training with synthesized noise signals

- CNN learns the required information for DOA estimation from the phase map
- CNN can be trained using synthesized noise signals (!?)
- Advantages:
  1. No speech/audio database required
  2. Easier to create training data
- Spectrally white noise was used

System overview



DOA estimate

$$\hat{\theta}_n = \underset{\theta_i}{\arg\max}\ p(\theta_i|\mathbf{\Phi}_n)$$

Evaluation
Experiments

1. Generalization to speech and robustness to noise
2. Performance in acoustic conditions different from training
3. Robustness to small perturbations in microphone positions
4. Real acoustic environments

Evaluation
Experiments

1. Generalization to speech and robustness to noise
2. Performance in acoustic conditions different from training
3. Robustness to small perturbations in microphone positions
4. Real acoustic environments

Evaluation
Performance measure

- Performance of CNN compared to SRP-PHAT
- Evaluation measure
  - Frame-level accuracy

$$A(\%) = \frac{\hat{N}_c}{N_s} \times 100,$$

$$N_s - \text{Total time frames with speech active}$$

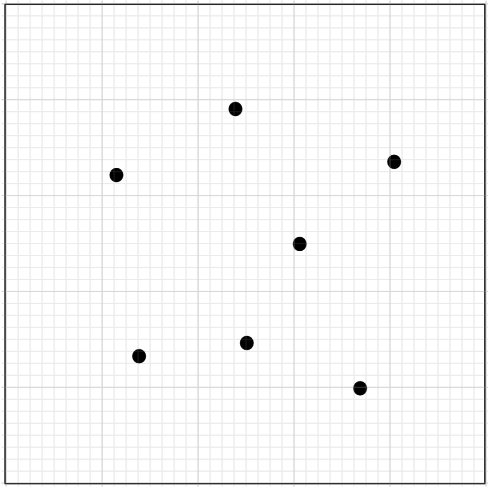
$$\hat{N}_c - \text{Time frames with correct estimate}$$

Evaluation
Experimental parameters

- Uniform linear array (ULA)
  - Number of microphones = 4
  - Inter-microphone distance = 3 cm
- STFT length 256, 50% overlap
- Resolution for classes: 5 degrees, I = 37 classes
- Training data is simulated using RIR generator [3]

[3] https://github.com/ehabets/RIR-Generator

Evaluation
CNN training conditions



Room with training positions

Simulated training data	
Signal	Synthesized noise signals
Room size	R1: (6×6) m , R2: (5×5) m
Array positions in room	7 different positions in each room
Source-array distance	1 m and 2 m for position
RT_{60}	R1: 0.3 s, R2: 0.2 s
SNR	Uniformly sampled from 0 to 20 dB

Evaluation

CNN training parameters

- **Training data**: 5.6 million time frames
- **Validation data**: 20% split from the training data
- **Loss**: Cross-entropy
- **Activation**: ReLU, Softmax (final layer)
- **Optimizer**: Adam
- **Batchsize**: 512
- **Nb Epochs**: 10
- **Regularization**: Dropout rate 0.5 (After Conv.3 layer, and each FC layer)

Evaluation
Test conditions

- **Database**: TIMIT test
- **Speech samples**: 500, 4 s each
- **Test data size**: 100000 active time frames

Simulated test data	
Signal	Speech signals from TIMIT
Room size	Room 1: (7×6) m , Room 2: (8×8) m
Array positions in room	1 random position for each room
Source-array distance	1.5 m for both rooms
RT_{60}	Room 1: 0.45 s, Room 2: 0.53 s
SNR	2 categories: 5 dB, and 15 dB

Evaluation

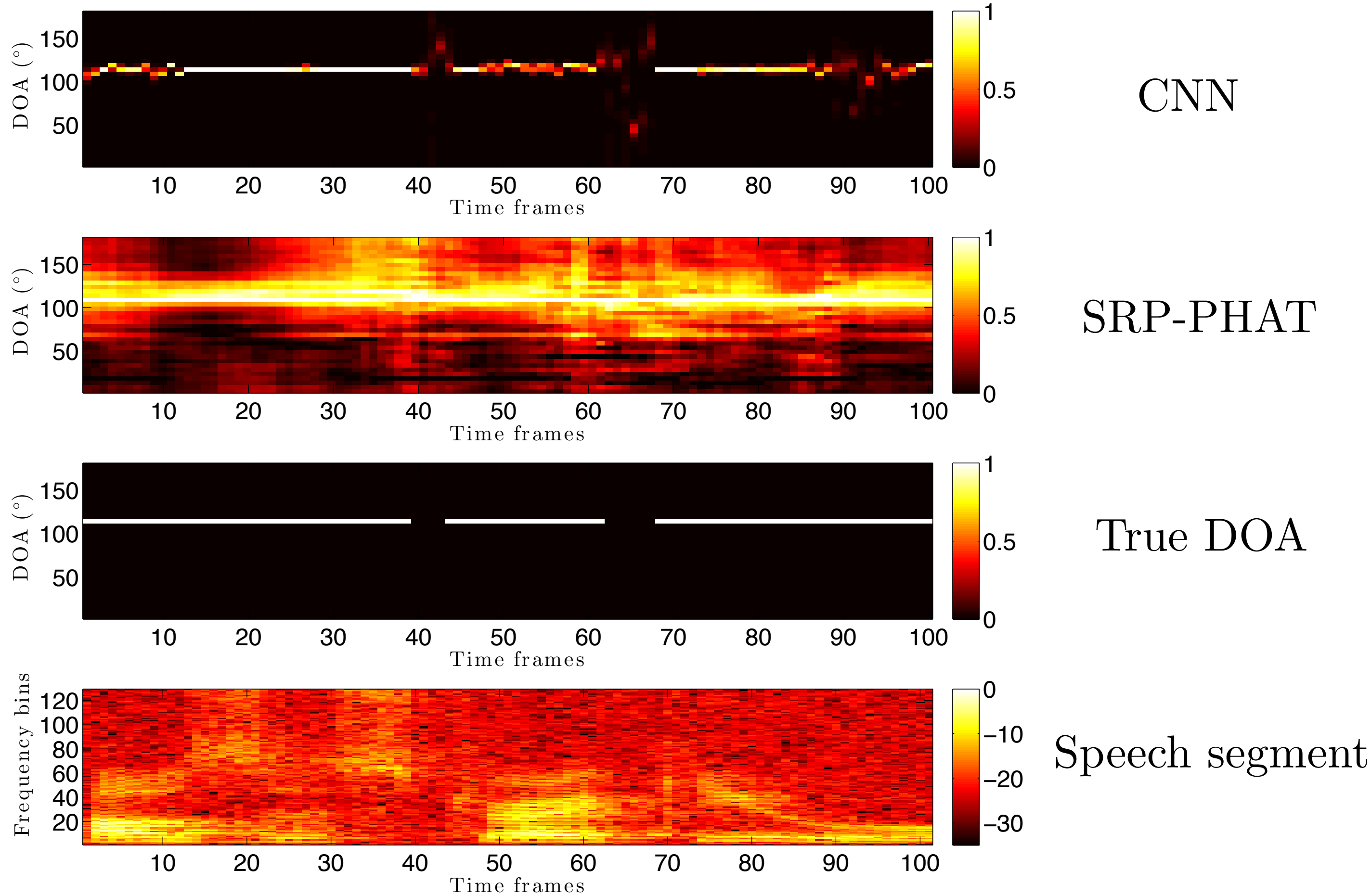
Test conditions

Frame-level accuracies (%)

	Room 1		Room 2	
	5 dB	15 dB	5 dB	15 dB
CNN	56.2	69.8	54.1	68.2
SRP-PHAT	22.6	33.6	21.8	38.4

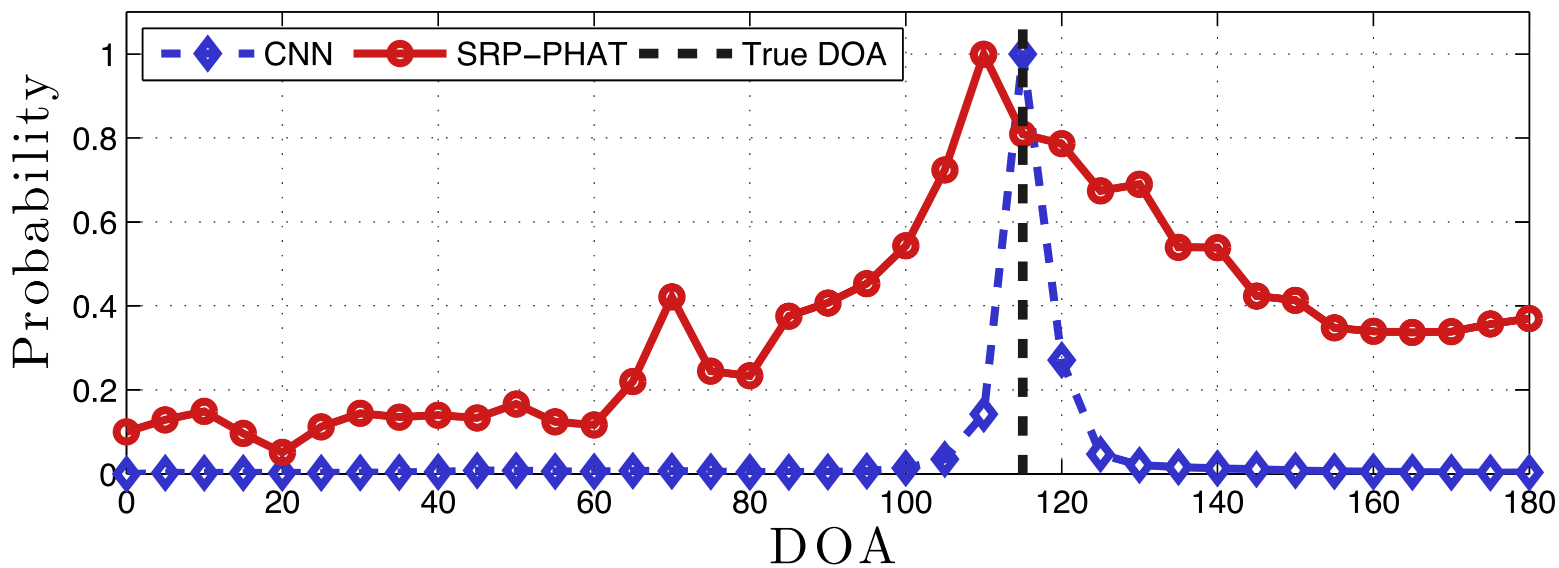
- Better performance compared to SRP-PHAT
- For all cases, CNN is accurate for the majority of the frames

Evaluation
Qualitative results - Room 2, 15 dB



Evaluation

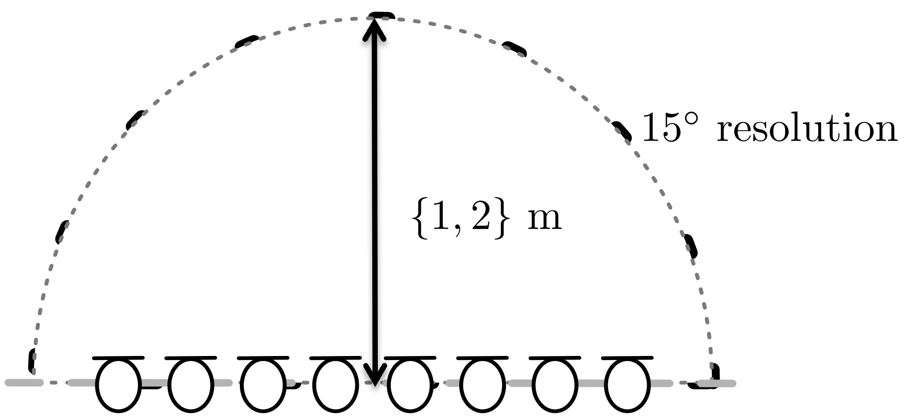
Qualitative results – Average over 0.8 s



Evaluation

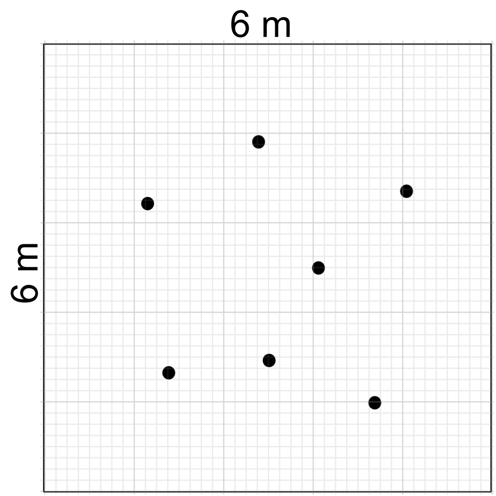
Real environment – Measured RIRs

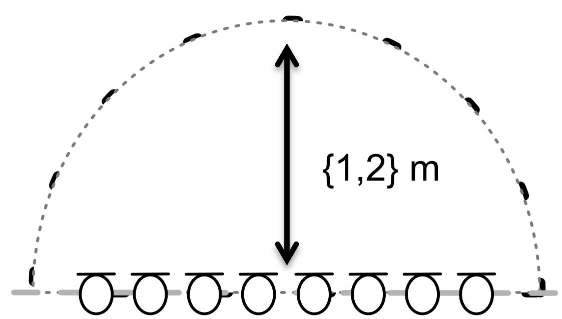
- **Database**: Multichannel Impulse Response Database from Bar-Ilan
- **Source setup**: $[0^\circ, 180^\circ]$, steps of 15 degrees
- **Array configuration**: M = 8 microphones, d = 8 cm
- **Source-array distances**: 1 m and 2 m
- **Test sample**: 15 s long speech sample



Evaluation
Real environment – Measured RIRs

- **Database**: Multichannel Impulse Response Database from Bar-Ilan
- **Source setup**: $[0^\circ, 180^\circ]$, steps of 15 degrees
- **Array configuration**: [8,8,8,8,8,8,8], M = 8 microphones
- **Source-array distances**: 1 m and 2 m
- **Test sample**: 15 s long speech sample
- CNN was retrained for the new array geometry (with simulated data)





Evaluation
Real environment – Measured RIRs

Frame-level accuracies (%)

	$RT_{60} = 0.160$ s		$RT_{60} = 0.360$ s		$RT_{60} = 0.610$ s	
	1 m	2 m	1 m	2 m	1 m	2 m
CNN	91.8	88.7	86.8	79.4	72.3	67.3
SRP-PHAT	94.4	69.0	87.1	68.3	71.7	62.4

- For 2 m distance, CNN outperforms SRP-PHAT
- SRP-PHAT performs better at lower reverberation times for 1 m source-array distance

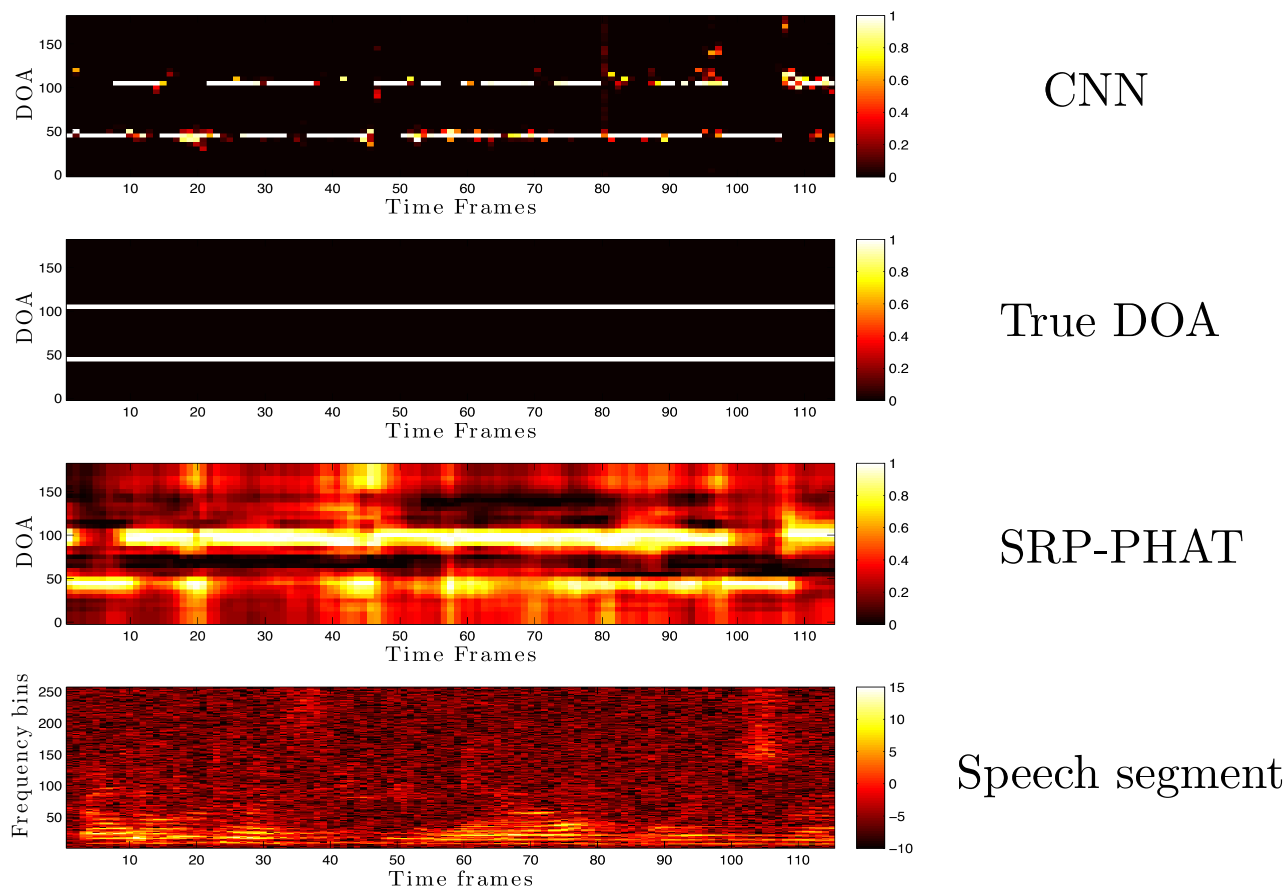
Conclusions

- Proposed a CNN based supervised learning method for DOA estimation with a simple input representation
- CNN trained with synthesized noise signals is able to localize a speech source
- Proposed system performs better than SRP-PHAT in unmatched simulated acoustic conditions
- Adaptability to unseen real acoustic environments was also demonstrated

Current Work
Multi-speaker localization

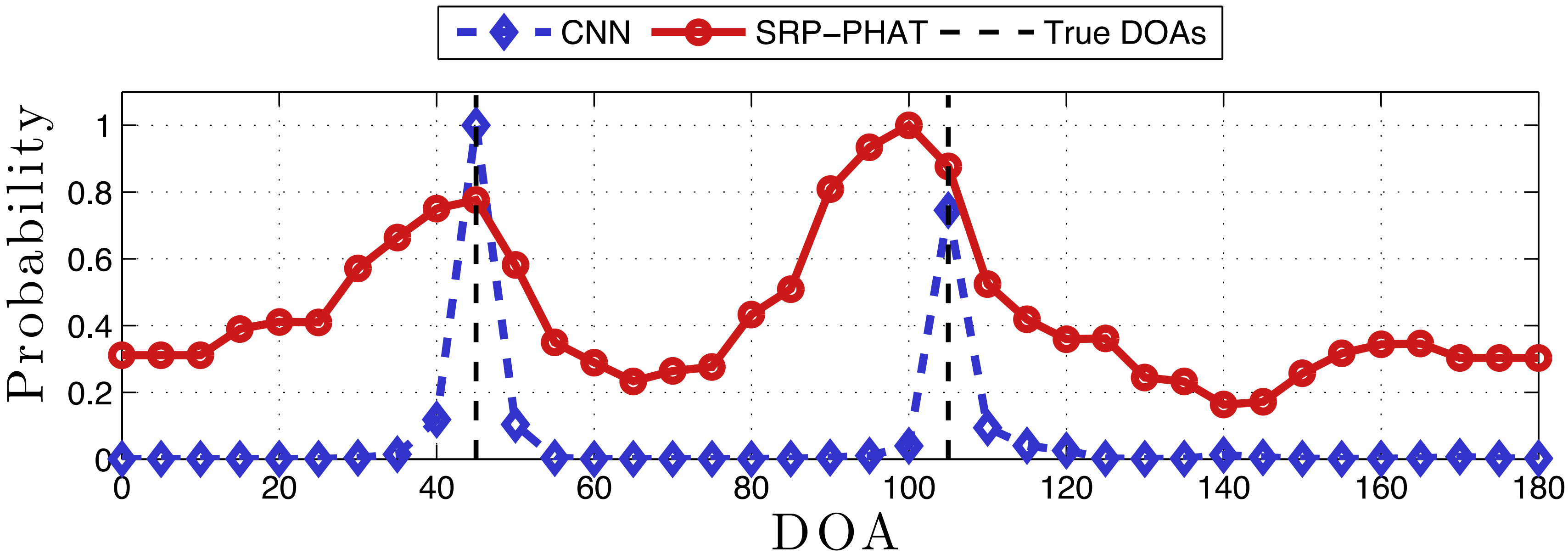
- Aim: Localize simultaneously active speakers
- Formulate multi-speaker localization as a multi-class multi-label classification problem

Current Work
Multi-speaker localization



Current Work

Multi-speaker localization



Still trained with synthesized noise signals!!

Thank you for your attention!

Trained model and weights available:

Soumitro-Chakrabarty/Single-speaker-localization