PAU
Perception and Activity Understanding Group

Deep learning for robust audio perception in human-robot interactions

Jean-Marc Odobez

IDIAP/EPFL Senior researcher – Head of PAU group

2019 ICRA workshop on

Sound Source Localization and its Applications for Robots

MuMMER
mummer-project.eu

Switzerland

EPFL, Lausanne

Zurich

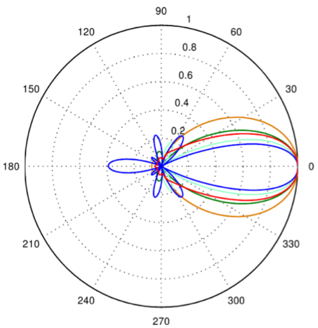
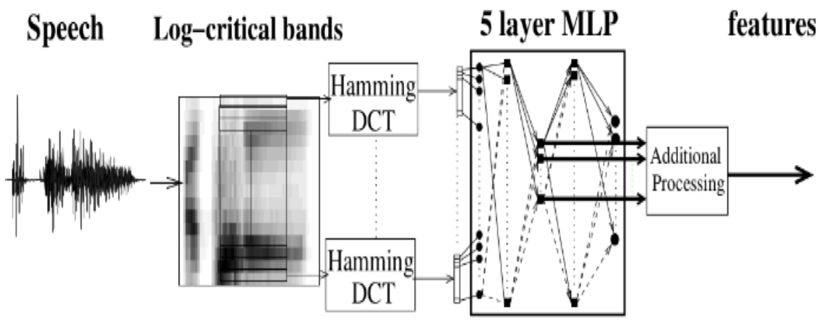
Geneva

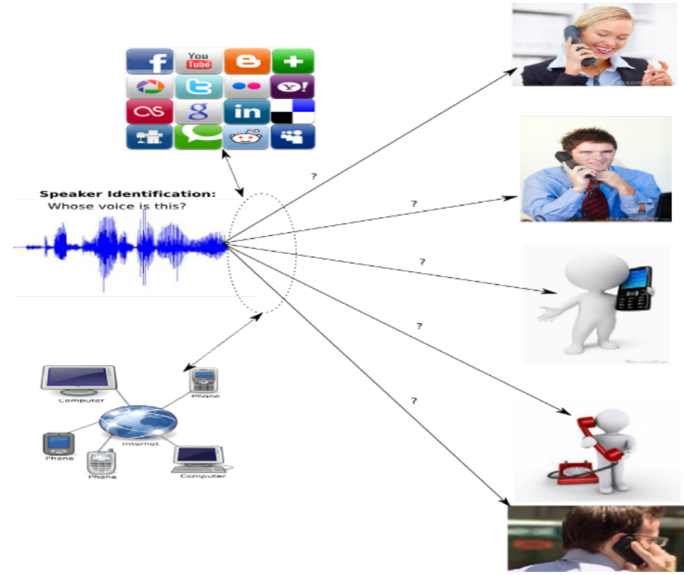
Idiap Research Institute (Martigny)

IDIAP Research Institute, in (very) brief

- Non for profit Foundation, created in 1991
  - academic affiliation with EPFL
- Human resources: around 100 people
  - 14 permanent researchers + 50 research associates (postdocs, PhD students) from more than 25 countries
- Main research areas - Artificial Intelligence for society
  - Machine Learning
  - Perceptual and Cognitive systems (speech, computer vision, natural language processing)
  - Human and Social Behavior (face-to-face communication, mobile, social media analysis,…)
  - Biometry
  - Robotics
  - …..

Speech group at Idiap



- Head: Prof. Hervé Bourlard
  - Researchers: Petr Motlicek, M. Magimai-Doss, Phil Garner
  - 25+ persons (researchers, phds, postdocs, interns, ...)
- Most speech related tasks
  - Forefront of Automatic Speech Recognition -- multilinguality
  - Speaker analysis (verification, identification, diarisation, role detection)
  - Microphone arrays and localization (beamforming, ad-hoc architectures)
  - Text-to-speech synthesis
  - Pathological speech processing
  - Speech assessment

PAU group research themes and objectives

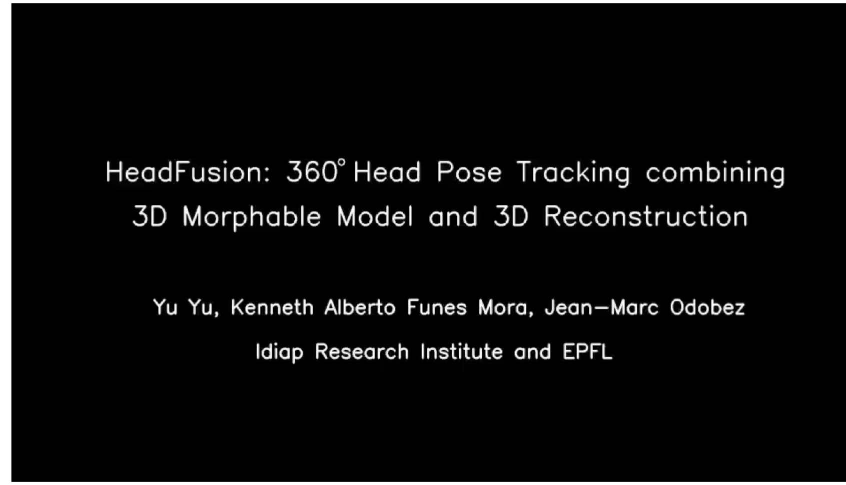
Thematic : sensing, interpreting, understanding

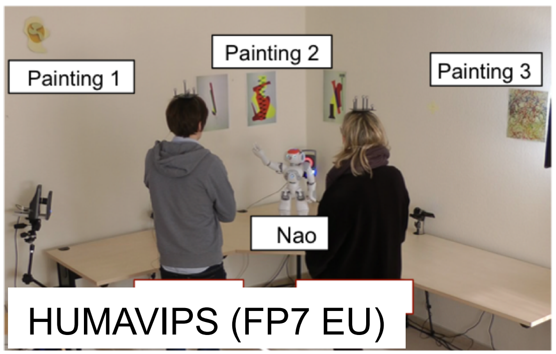
- Perceptual component
  extraction physical representations - detection, tracking, pose
- Activity understanding
  gestures, behaviors - individual, group level - context
- Methods & Models
  computer vision, (multimodal) signal processing, sociology
  machine learning: statistical models; deep learning
- Applications
  surveillance, human-robot interfaces, sociology, multimedia content analysis

Interaction analysis


Idiap - KTH dataset

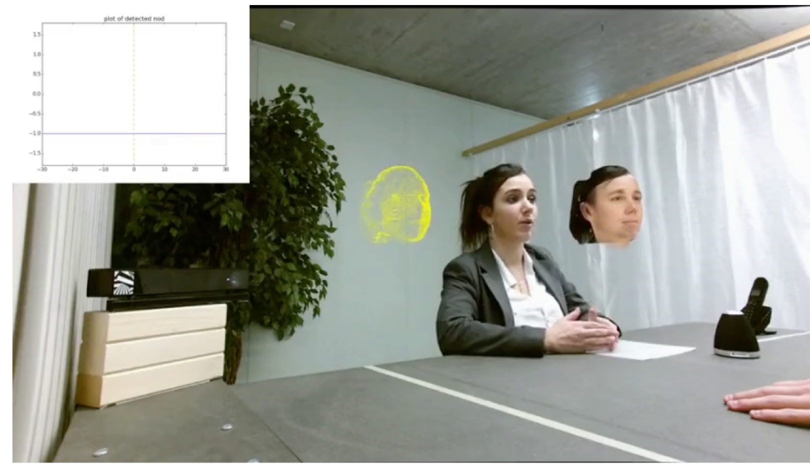
HeadFusion
360 degree head tracking





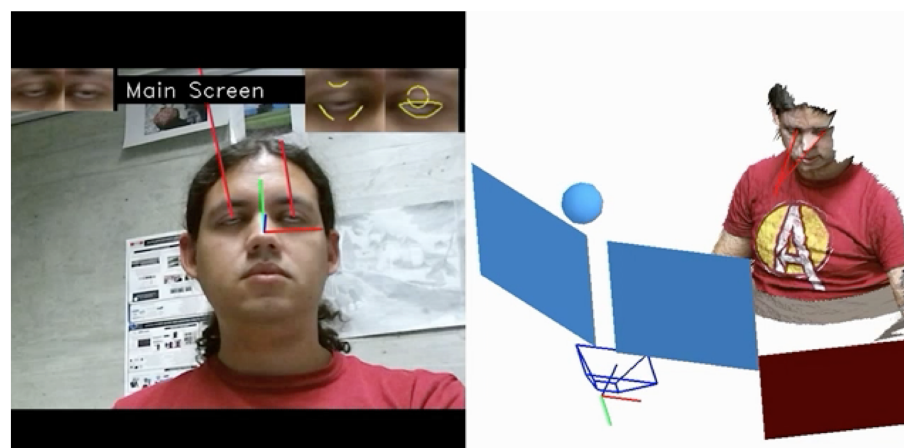
HUMAVIPS (FP7 EU)

Head gestures (nods)



MUMMER (H2020 EU)

Gaze & attention

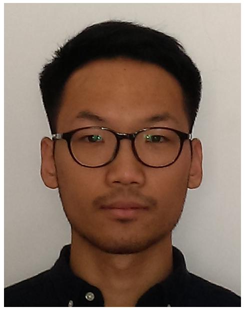


- Set-ups & Tasks (tracking, re-id, non-verbal cues)

PAU
Perception and Activity
Understanding Group

Deep learning for robust audio perception in human-robot interactions

Jean-Marc Odobez



Joint work with
- Weipeng He (Phd student)
- Petr Motlicek (researcher)

Outline

- Joint sound source localization and discrimination   with deep learning
- Multiple Sound source localization NN adaptation using weak labels

Interacting with robots : MuMMER EU project



- GOAL: Develop a humanoid robot
  - public shopping mall
  - entertaining, give information, directions
  - autonomous, natural interactions

- Participants
  - University of Glasgow (UK)
  - Heriott-Watt University (UK)
  - Idiap Research Institute (CH)
  - LAAS-CNRS (France)
  - Softbank Robotics Europe (France)
  - VTT Technical Research Center (Finland)
  - Ideapark (Finland)

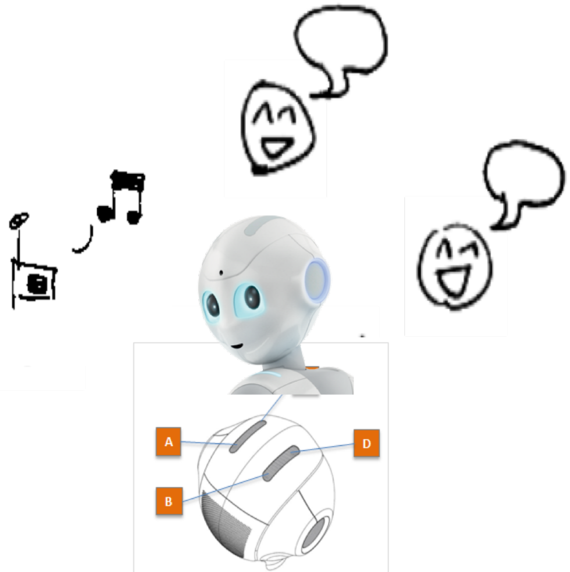
Sound source localization & discrimination

MUMMER (H2020 EU)

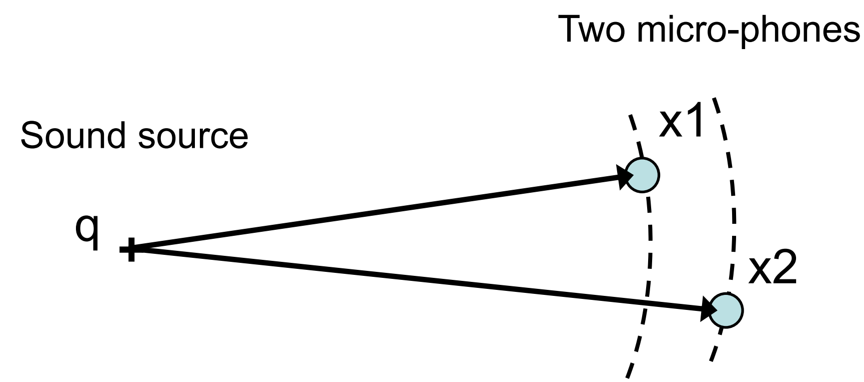
- Challenges
  - Unknown number of sound sources
  - Strong noise (robot ego-noise, background)

  - Speech and non-speech sources
  - Speech overlap (simultaneous speakers)
  - Short utterances during interactions

Sound source localization



Micro-phone array



$$R_{12}(\tau) = \frac{1}{2\pi} \int \Psi_{12}(\omega) X_1(\omega) X_2(\omega)^* e^{j\omega\tau} d\omega$$

GCC-PHAT

- Traditional approaches (localization) : signal processing
  - Interaural time and intensity differences
  - Time difference of arrival (TDOA)
    - E.G. :   GCC-PHAT: Generalized Cross-Correlation with Phase Transform

=> relies on modeling assumptions

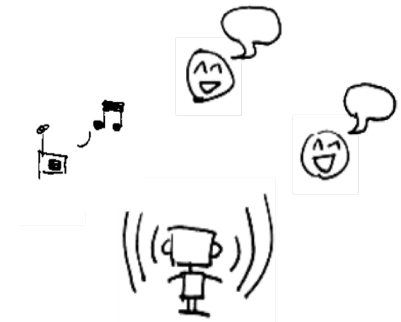
(head model, geometry knowledge, obstacles, propagation, …)

Sound source discrimination

Localization ⇨ Beamformer/ Separation ⇨ Classification

- Previous methods : solve the problem sequentially
- Issues:
  - beamforming : enhances the signal coming from a given direction
    - direction is approximately known
  - different audio representation/processing for localization and classification
    - related: which signal frequencies comes from which direction ?

Sound source localization & discrimination



Localization

Spatial information →

← Spectral information

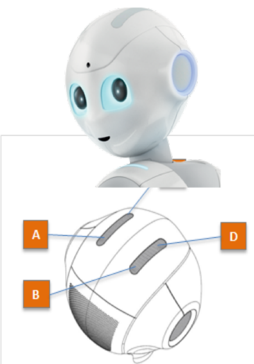
Classification

- Proposition: learning-based joint localization & discrimination
  - both tasks help each other
  - fewer assumption required
  - direct optimization for the tasks

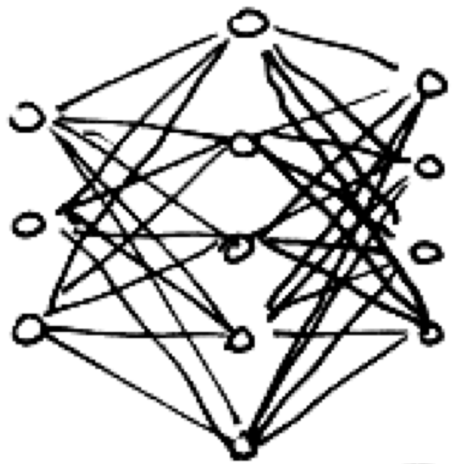
Deep Neural Networks for Multiple Speaker Detection and Localization, He, Motlicek, Odobez, Int. Conference on Robotics and Automation (ICRA) 2018

Joint Localization and Classification of Multiple Sound Sources Using a Multi-task Neural Network, He, Motlicek, Odobez, Interspeech 2018

Sound source localization & discrimination



Input



Output

SSL : Sound source Localization

SNS : Speech/Non speech discrimination

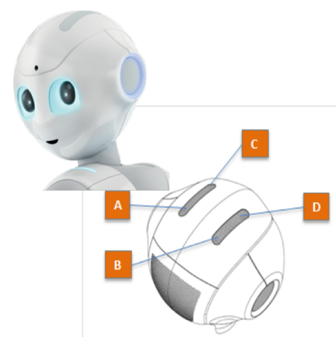
Architecture

Training

Training data

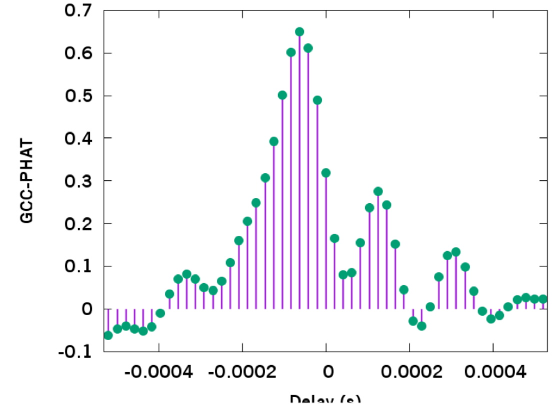
- How to proceed?

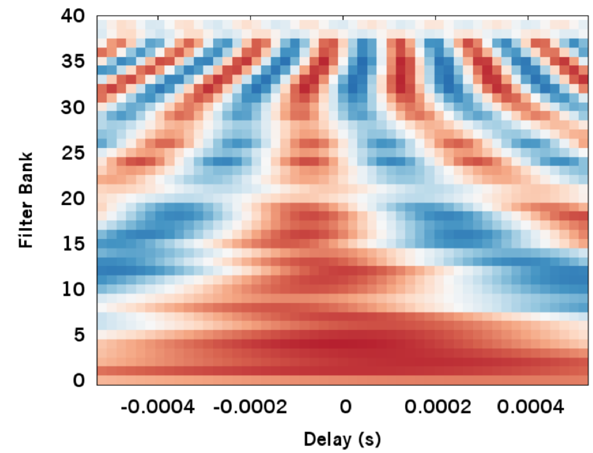
SSL & SNS : Input

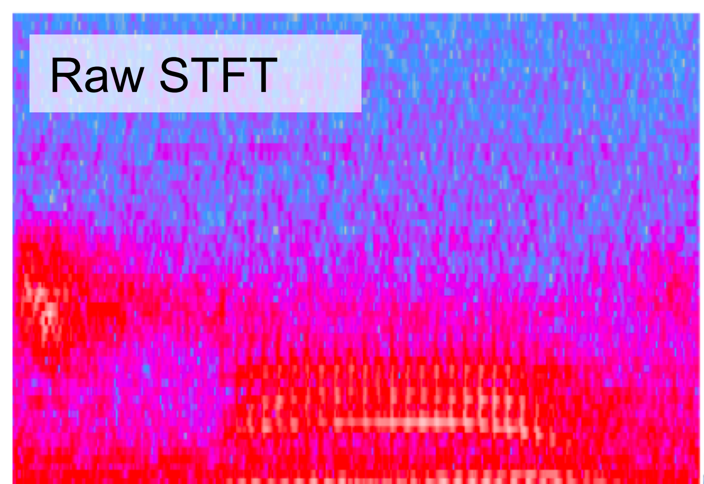


- Input signals
- Per **pair of** microphone (6)
  - GCC-PHAT delay coefficients
  - GCC-PHAT on filter banks
  - Ok for localization
  - Lacks spectral information for SNS

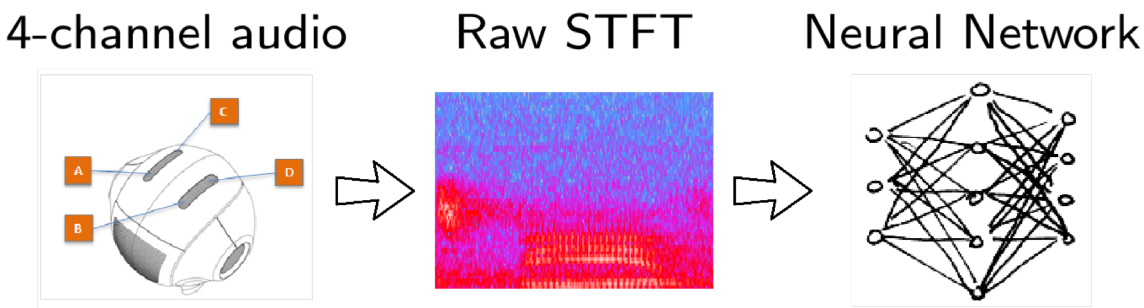
- Short-Time Fourier transform (SFTF) per microphone





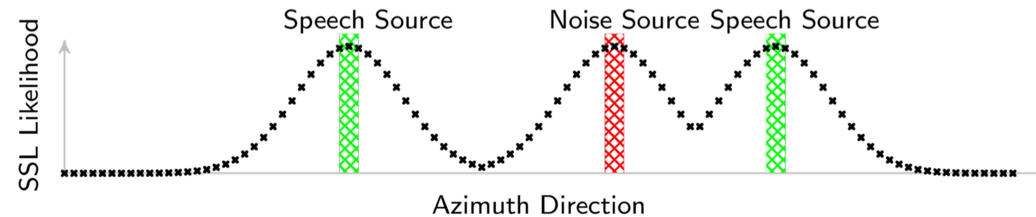


SSL & SNS : Output & Loss Function

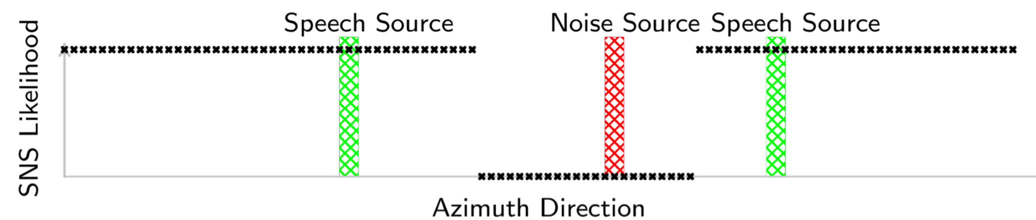


- Likelihood for each sound direction

$\boldsymbol{p}$:



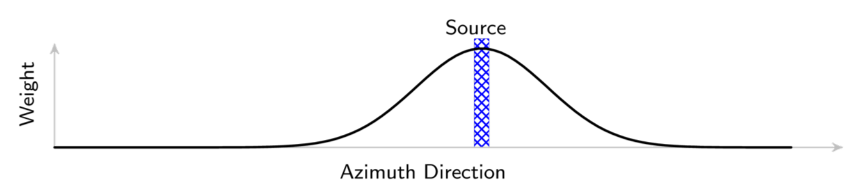
$\boldsymbol{q}$:



$$Loss = \|\hat{\boldsymbol{p}} - \boldsymbol{p}\|_2^2 + \sum_i \boxed{w_i} |\hat{q}_i - q_i|^2$$

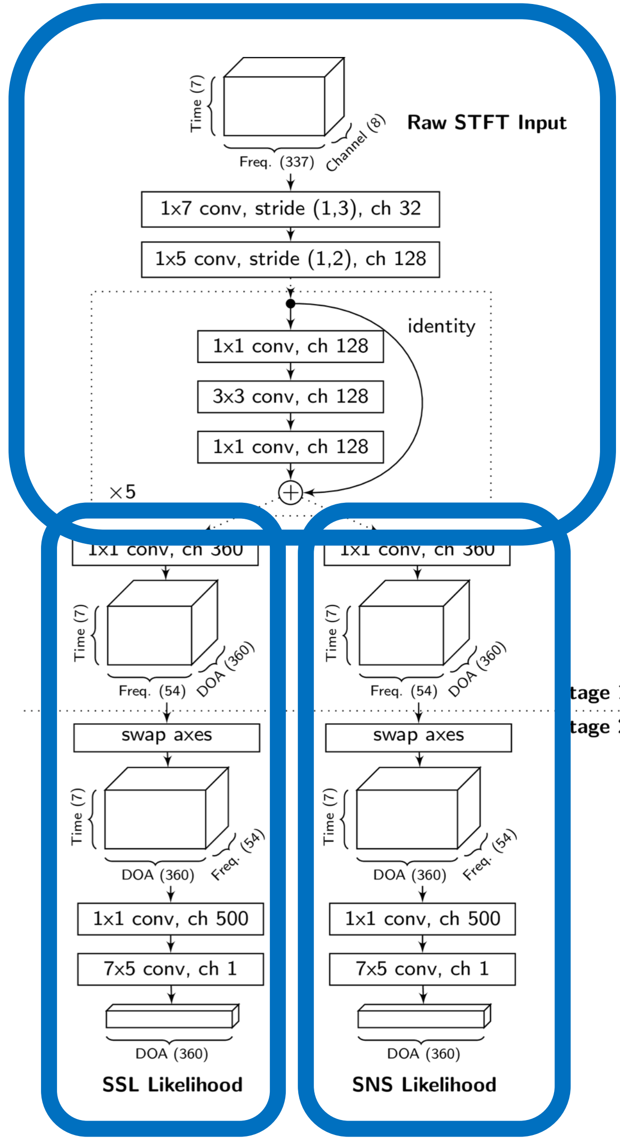
emphasis on directions next to active sources

$\boldsymbol{w}$:



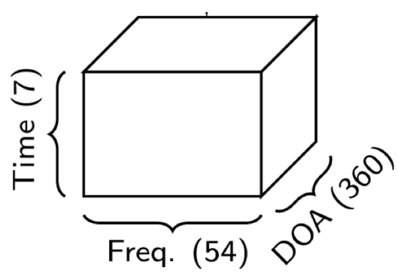
SSL & SNS : Network structure

- Fully convolutional
  - Residual network trunk
  - Two task-specific branches



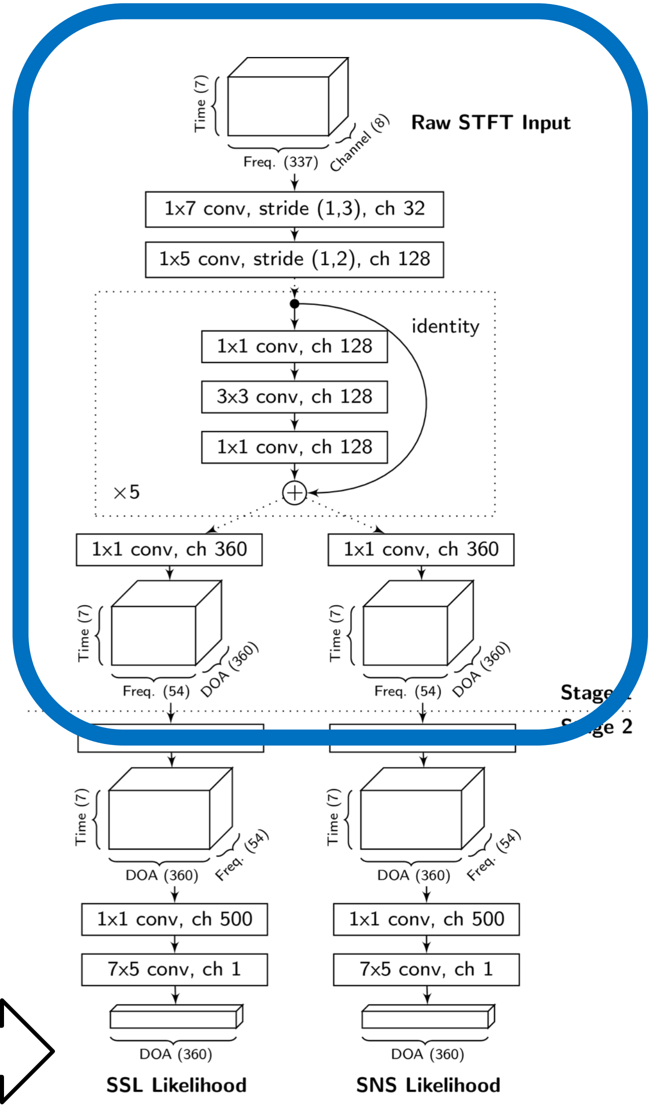
SSL & SNS : Network training

- End-to-end
  - Not working well
- Two-stage approach
  - Stage 1
    - Convolutions in Time-Frequency domain



    - Output: 360 (DOA) channels
    - early SSL & SNS prediction for each TF point

Train whole network end-to-end

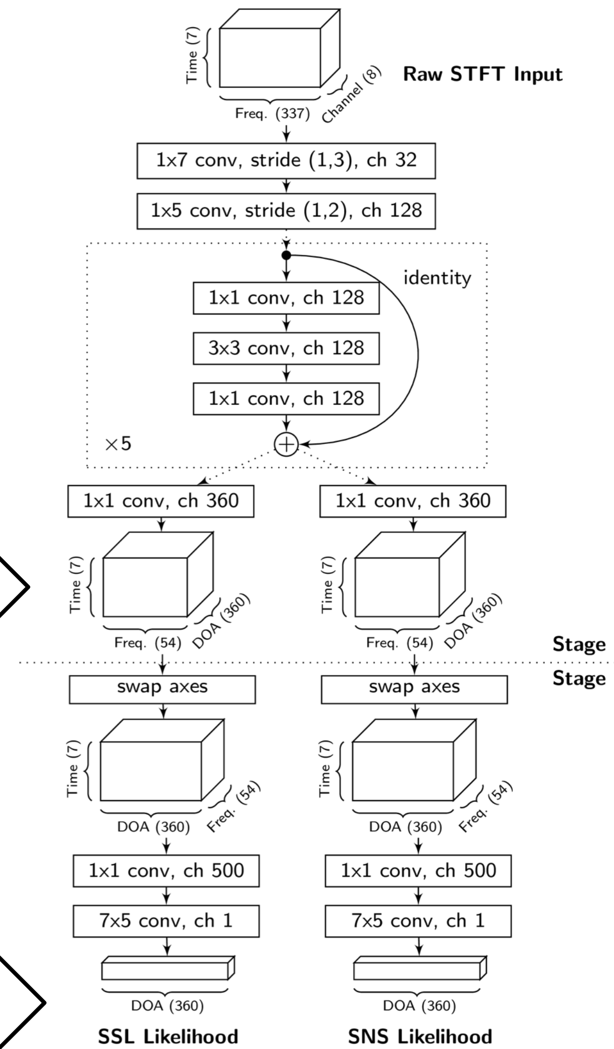


SSL & SNS : Network training

- End-to-end
  - Not working well
- Two-stage approach

(1) Supervision on output of stage 1

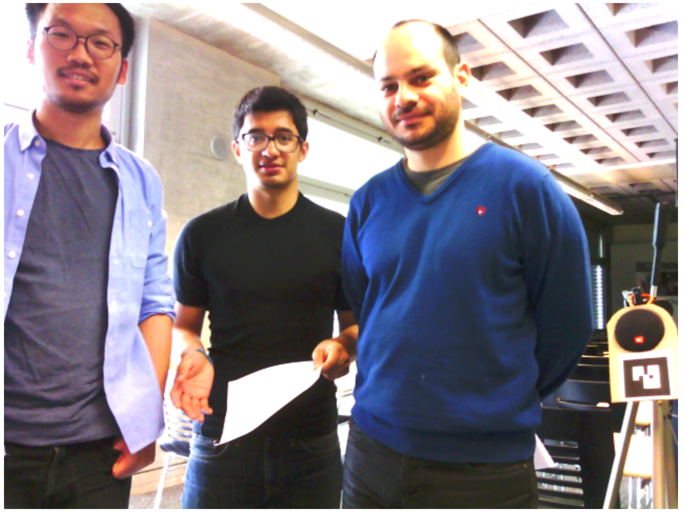
(2) Train whole network end-to-end



SSL & SNS : Experiments

- Training
  - Loudspeakers: 32 hours, 148 speakers
    - Speech: AMI Corpus
    - Non-speech: Google AudioSet
    - **Pepper moves to collect data with variabilities**
      => faster data-collection

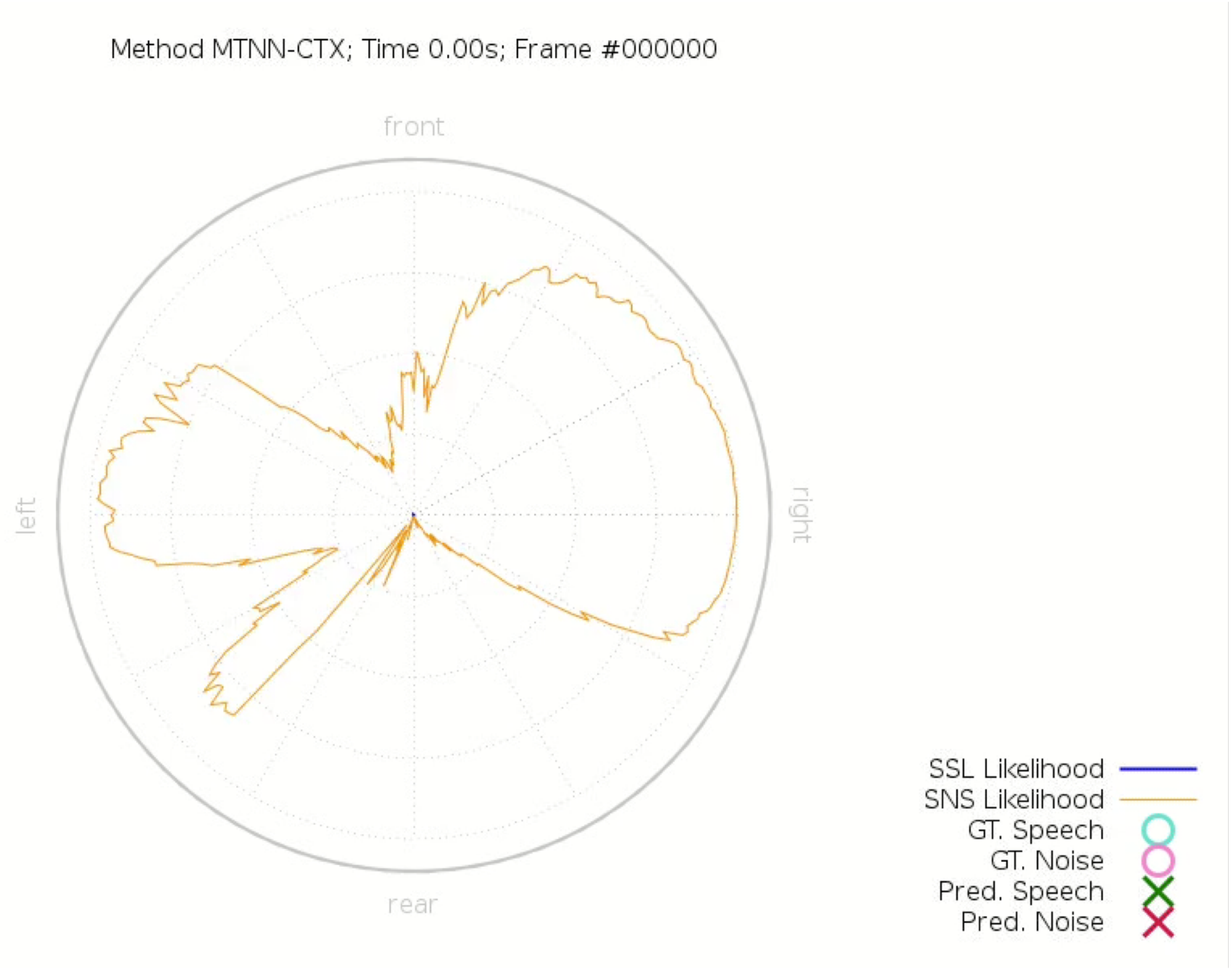
- Test
  - Loudspeaker: 17 hours, 16 (different) speakers
  - Human talkers: 8 minutes, 7 speakers
    (with loudspeakers Non Speech sources)



- Sound Source Localization for Robots (SSLR) Dataset
  https://www.idiap.ch/dataset/sslr

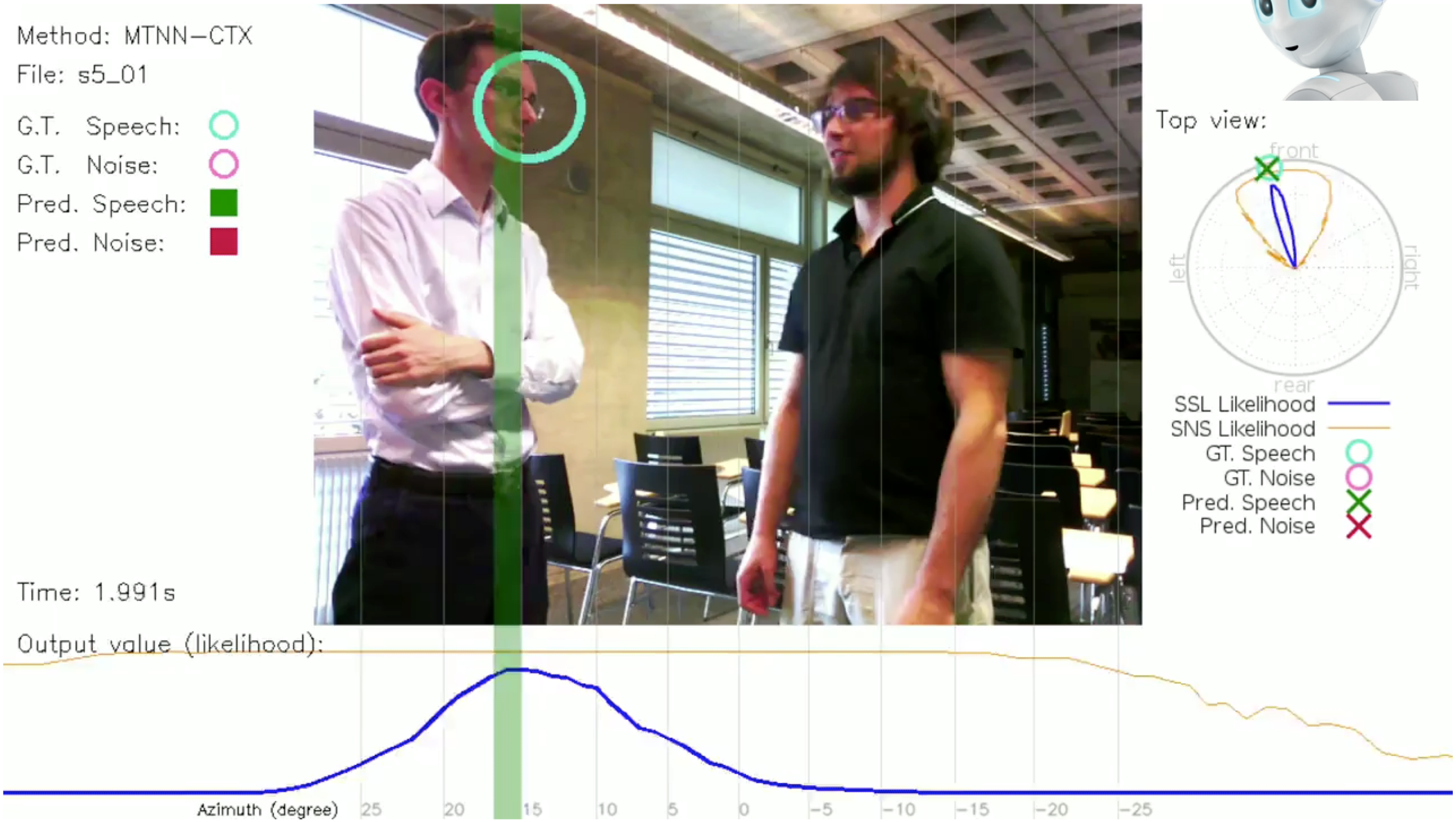
AMI Corpus: http://groups.inf.ed.ac.uk/ami/corpus/
AudioSet: https://research.google.com/audioset/

SSL & SNS : Experiments – qualitative results



- Loudspeaker recording

SSL & SNS : Experiments - qualitative



- Blue curve: likelihood of a sound source
- Yellow curve: is the source speech (1) or not (0)

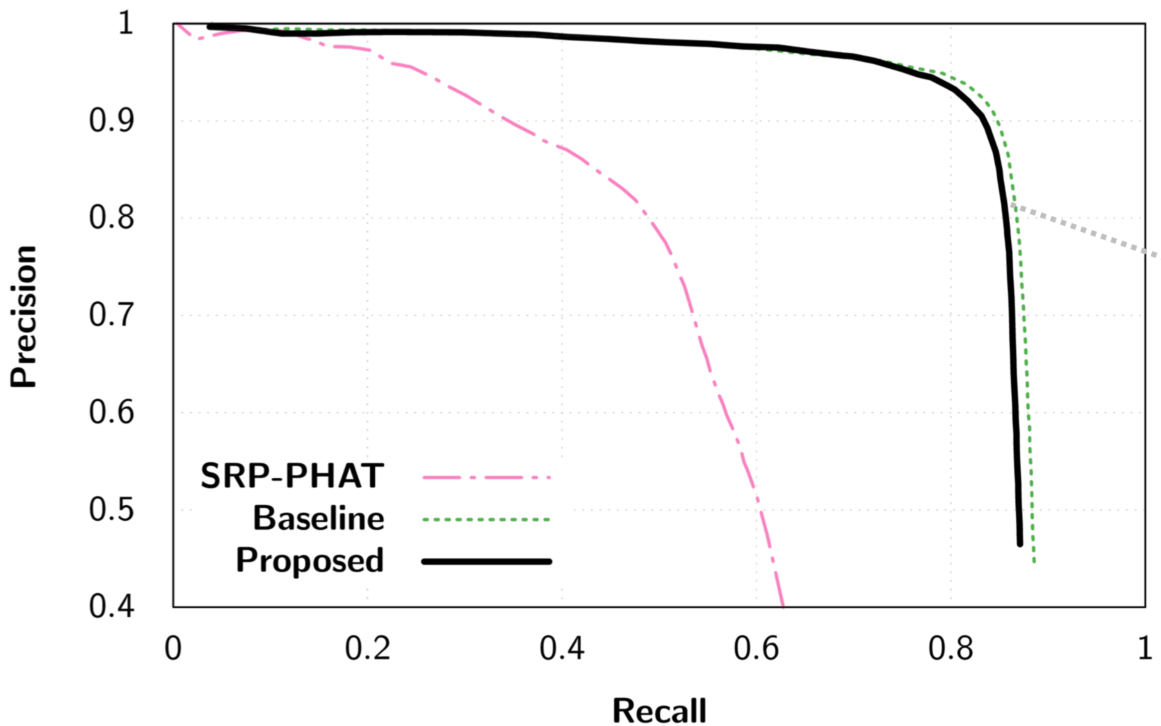
SSL & SNS : Experiments – evaluation

- Tasks
  - Sound Localization
  - Speech/Non-speech Classification
  - Speech Localization

- Methods
  - Baseline: two-step approach
    - localization NN
    - MVDR + classification (NN on beamformed signal)
  - Proposed method

SSL & SNS : Experiments – sound localization

- Human recordings



Proposed Method
>90% precision
>80% recall

(Similar conclusion can be drawn for loudspeaker recordings)

SSL & SNS : Experiments – speech/non-speech

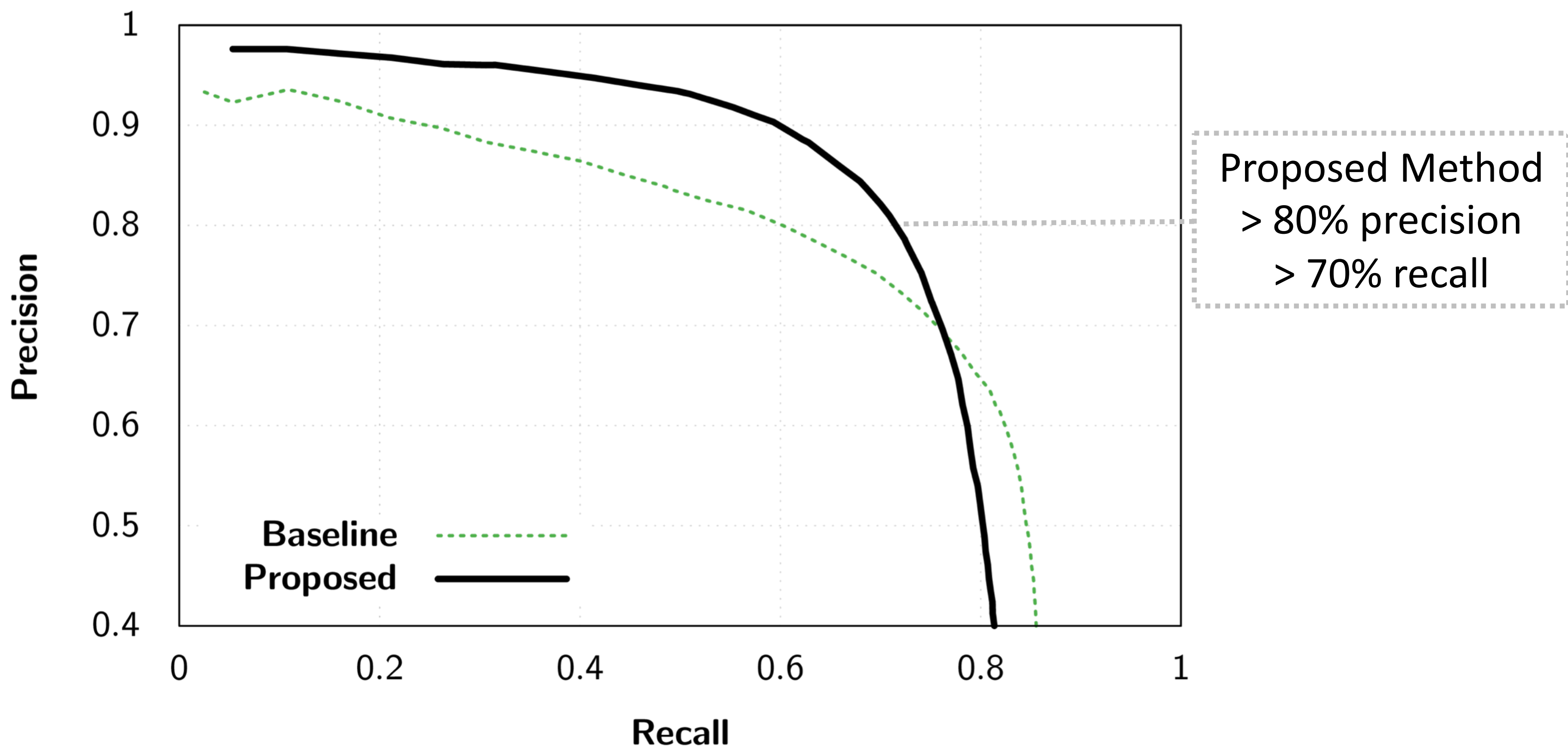
(Assuming sound direction is known)

Accuracy	Loudspeaker	Human
Baseline	0.80	0.68
Proposed Method	**0.95**	**0.85**

- Proposed method
  - much better
  - good generalization

SSL & SNS : Experiments – speech localization

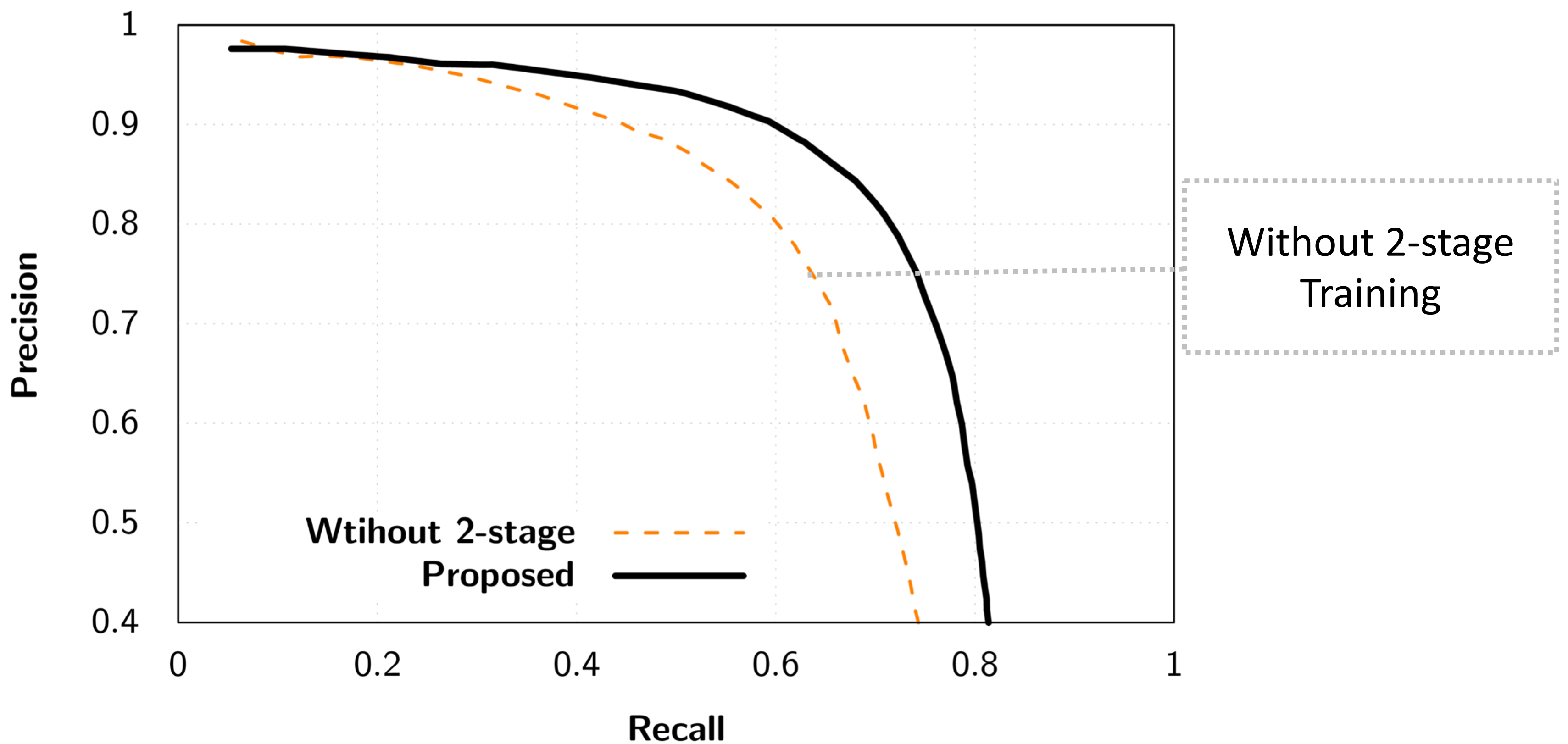
- Human recordings



(Similar conclusion can be drawn for loudspeaker recordings)

SSL & SNS : Experiments – speech localization

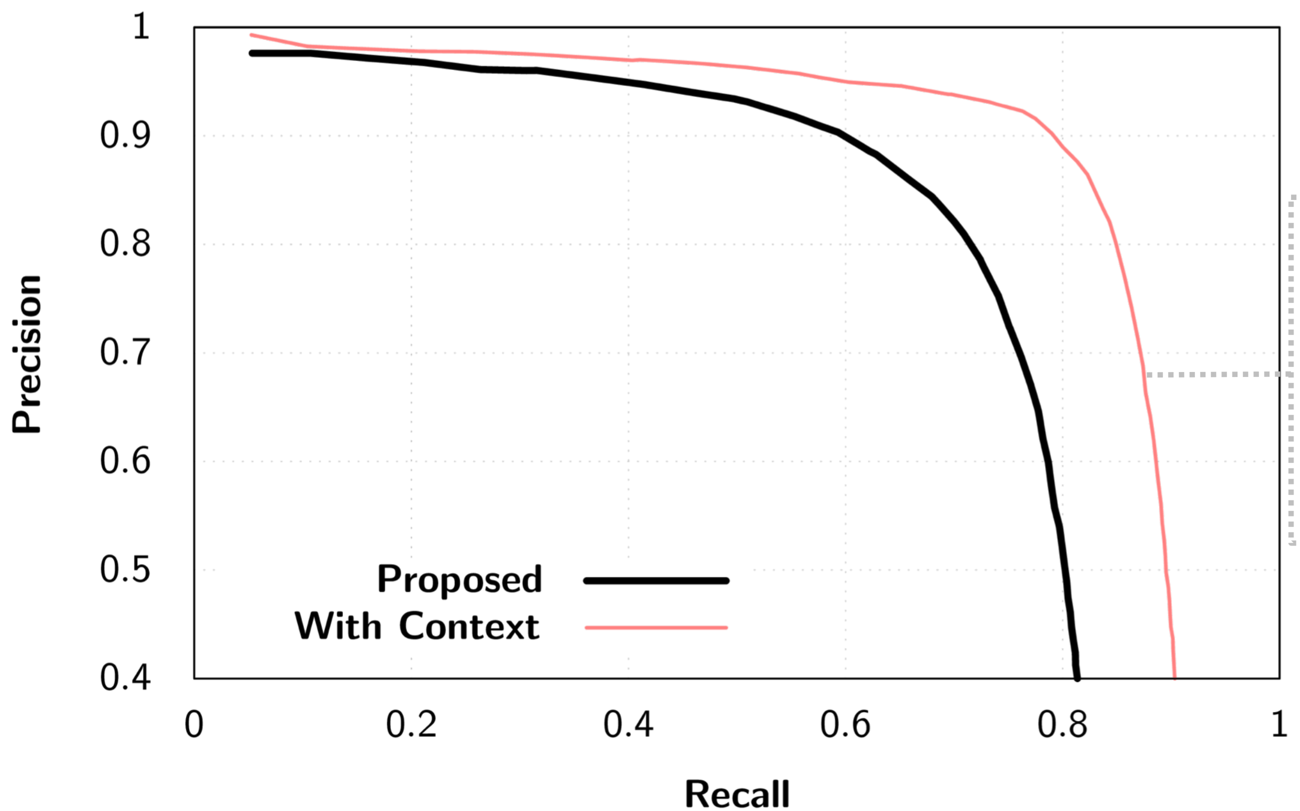
- Human recordings



(Similar conclusion can be drawn for loudspeaker recordings)

SSL & SNS : Experiments – speech localization

- Human recordings



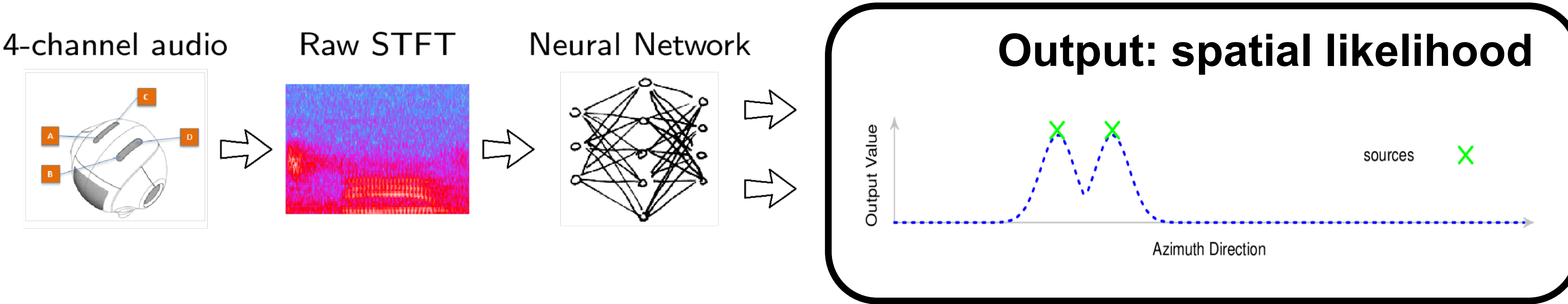
Extension:
Add Temporal Context to Input (600ms instead of 170ms)

(Similar conclusion can be drawn for loudspeaker recordings)

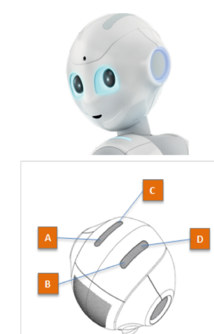
Outline

- Joint sound source localization and discrimination   with deep learning
- Multiple Sound source localization NN adaptation using weak labels

Learning to localize sound



- Current issue - training data
  - diversity in source signals (voices, noise), power, positions, noise, etc.
  - **device specific**
  - collection and annotation can be costly

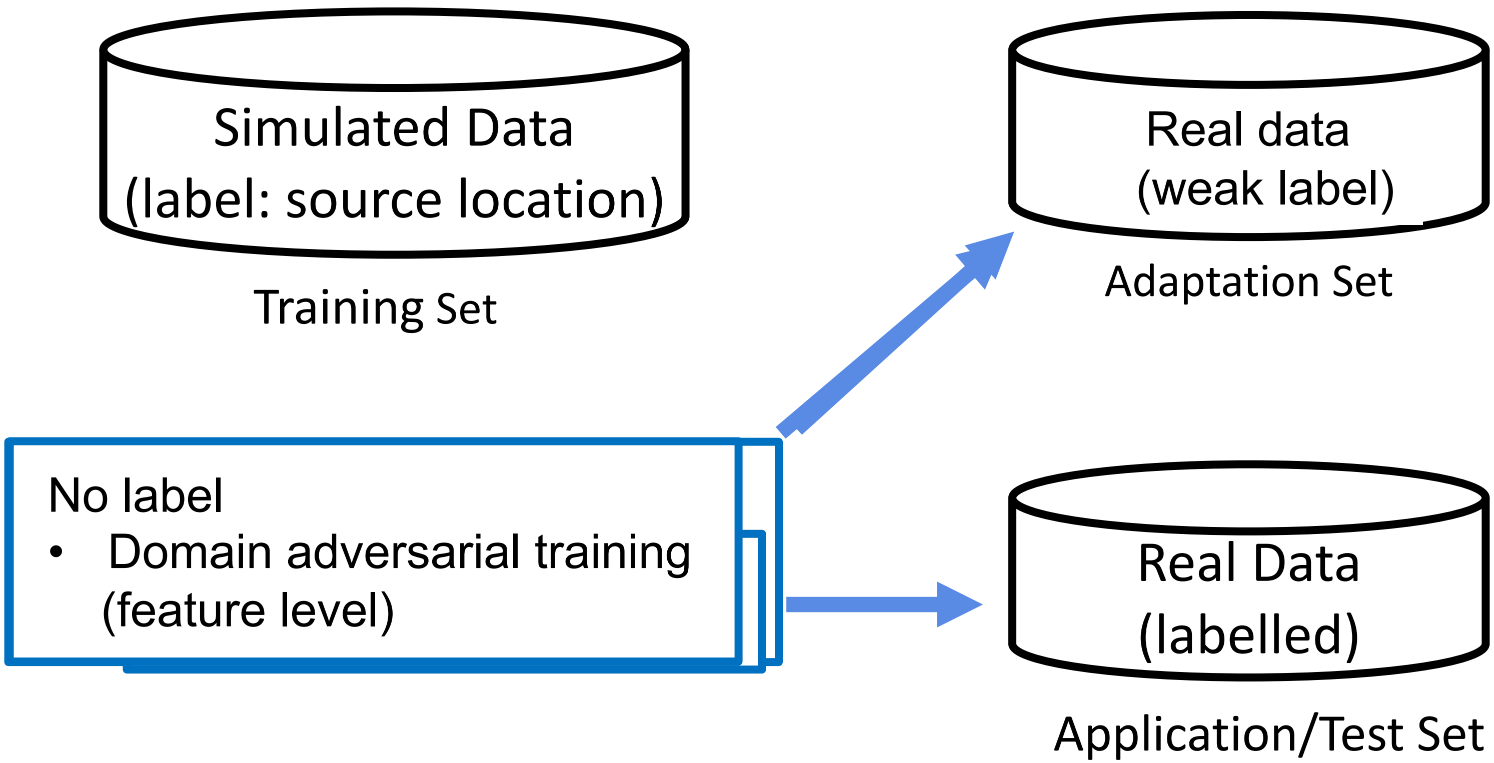


- Approach & motivations
  - train network **with simulated data –apply to real data?**
    - control diversity, exploit large datasets
  - **reality gap** : mismatch between simulation & real conditions
    - device physical body & microphone response pattern, room features,…

We need domain adaptation !

Domain adaptation – problem formulation

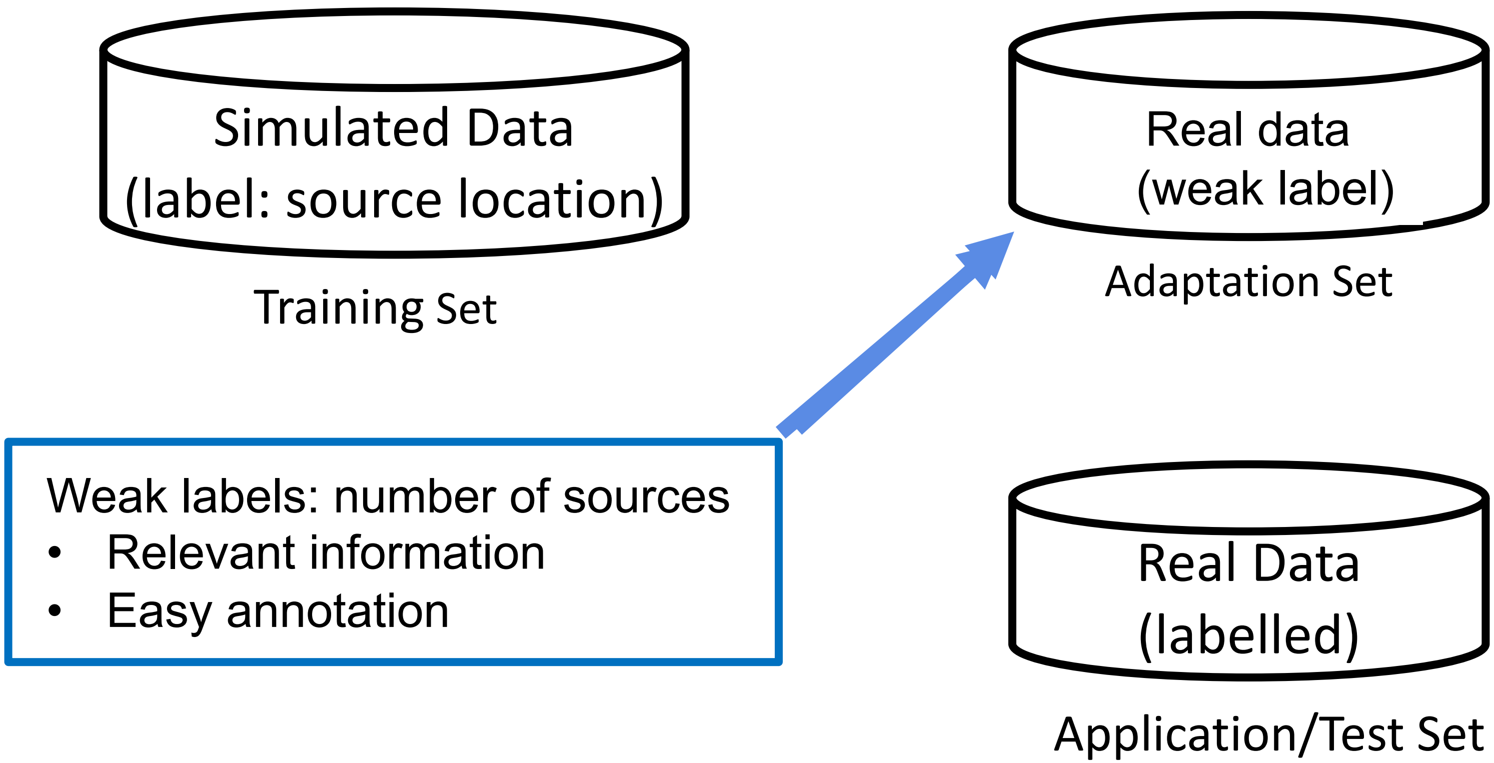
- Source domain
- Target domain



[Adaptation of Multiple Sound Source Localization Neural Networks with Weak Supervision and Domain-Adversarial Training](), He, Motlicek, Odobez, Int. Conference on Acoustic, Speech and Signal Processing (ICASSP) 2019

Domain adaptation – problem formulation

- Source domain
- Target domain

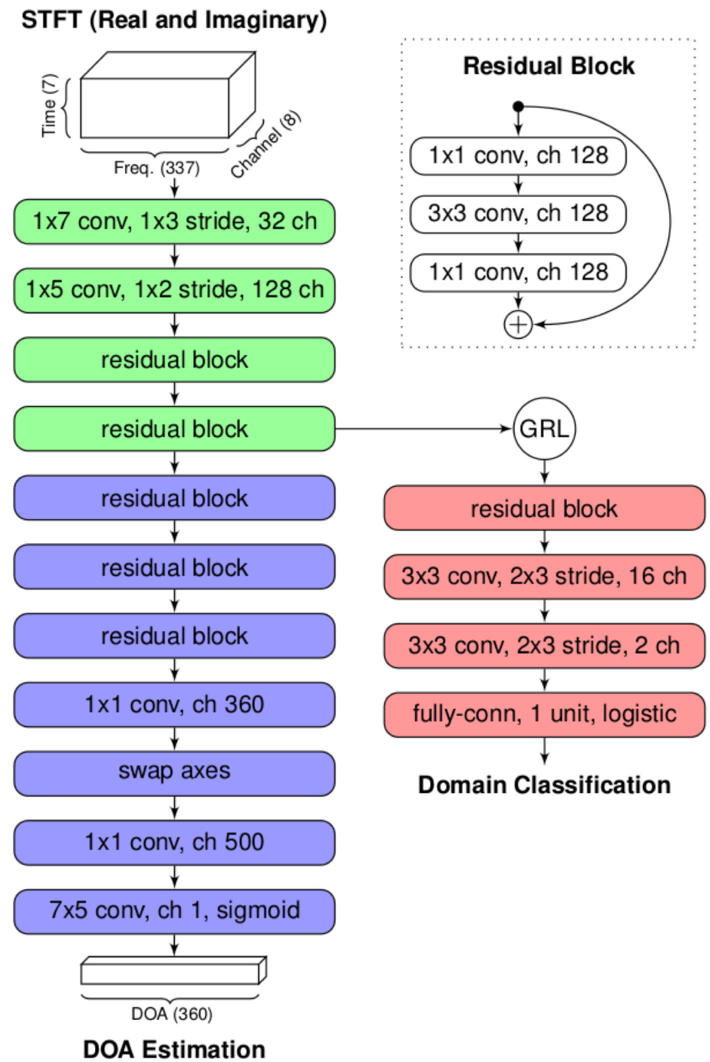


[Adaptation of Multiple Sound Source Localization Neural Networks with Weak Supervision and Domain-Adversarial Training](), He, Motlicek, Odobez, Int. Conference on Acoustic, Speech and Signal Processing (ICASSP) 2019

Domain adaptation – domain adversarial training

- Goal: learn features (green part)
  - Perform well for the task (purple part) (on simulated data)
  - Are domain independent (red part)

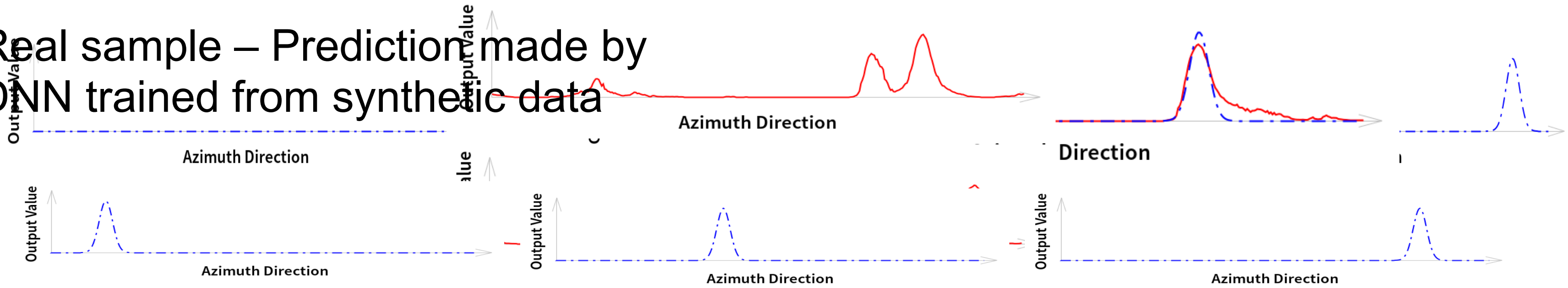
  => domain classifier can not distinguish those produced from the real or simulated data

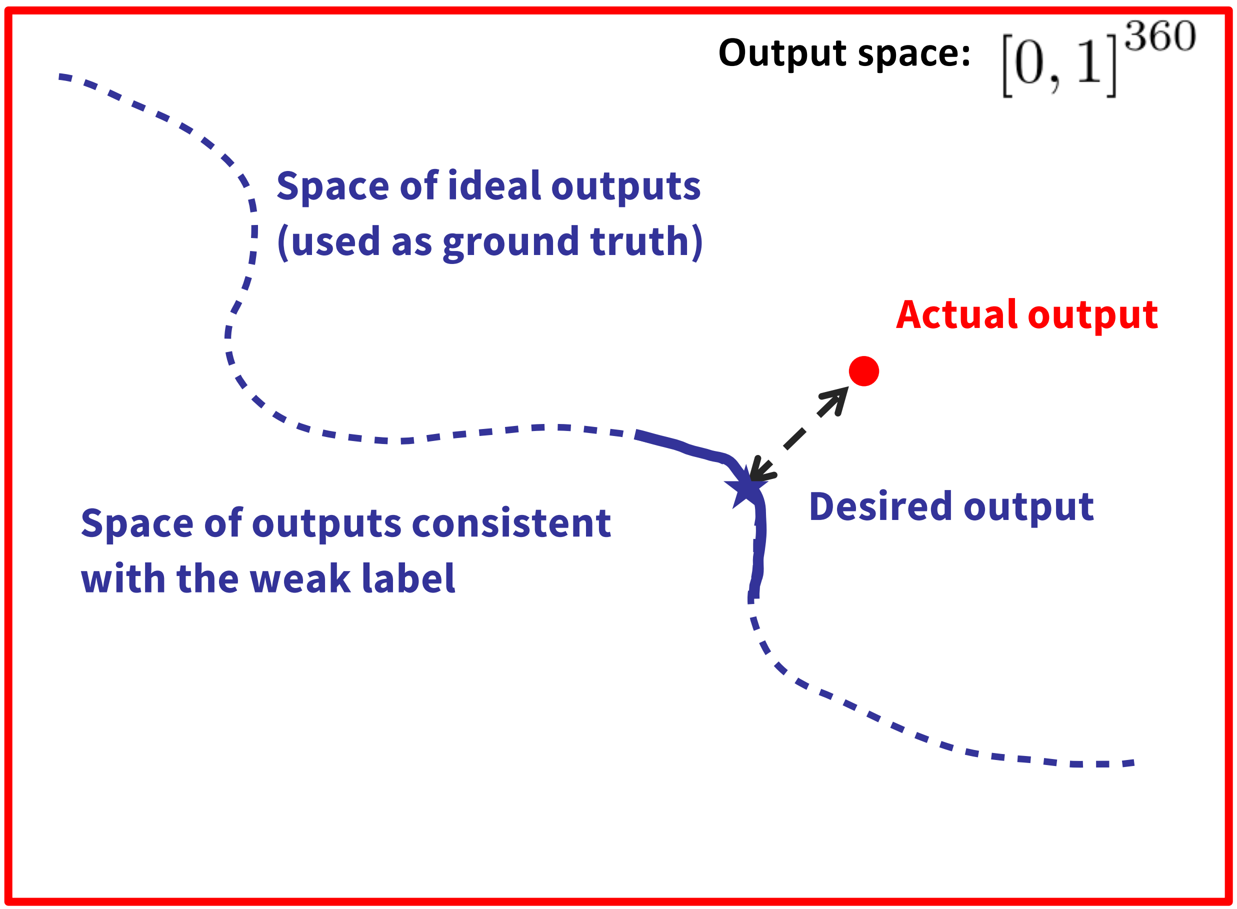


Ganin et al. "Domain-adversarial training of neural networks," Journal of Machine Learning Research, 2016.

Domain adaptation – weak supervision

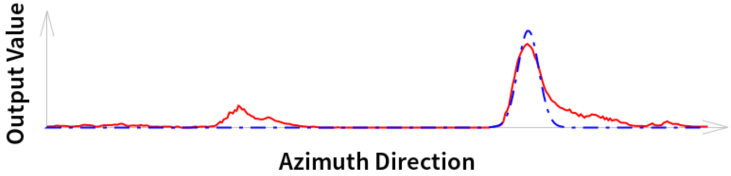
Real sample – Prediction made by DNN trained from synthetic data

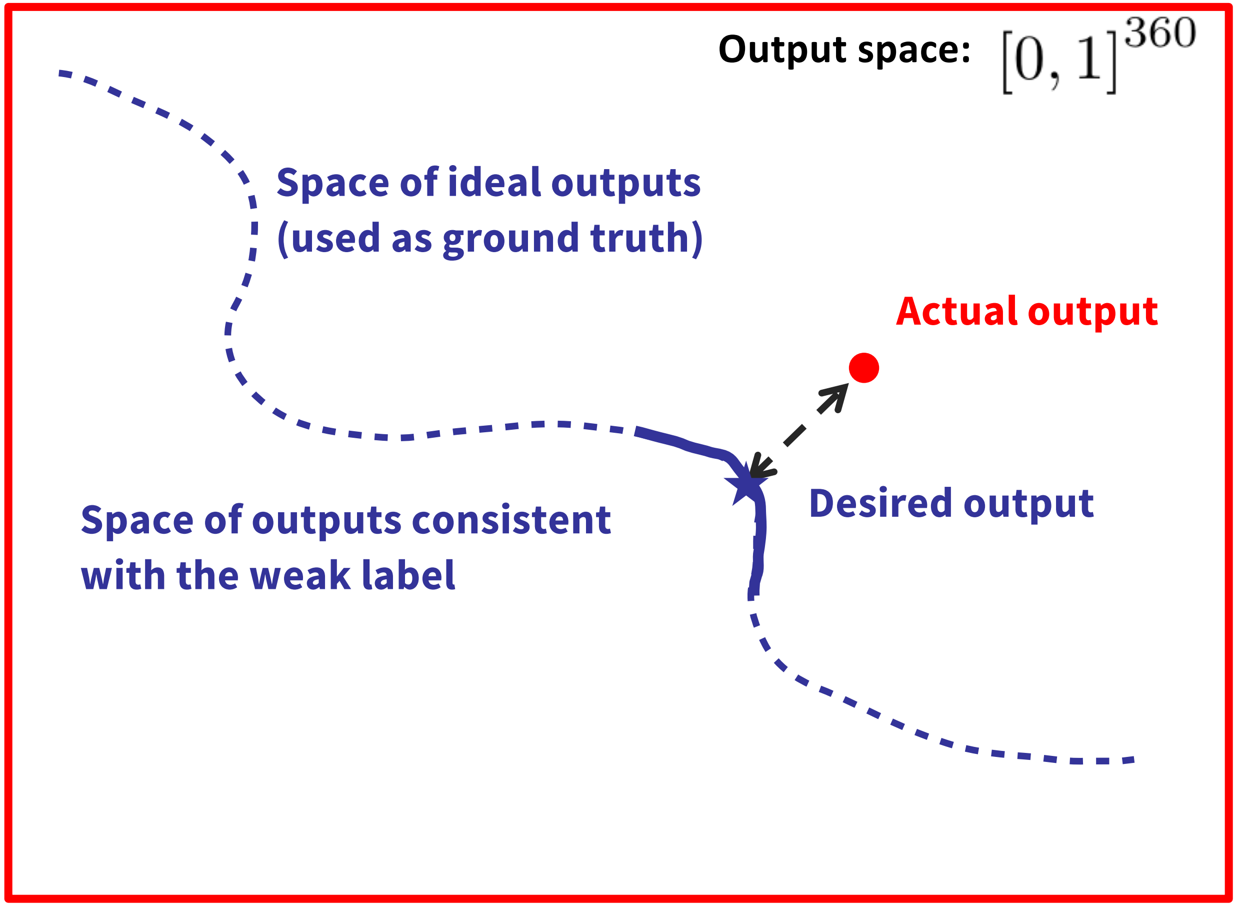




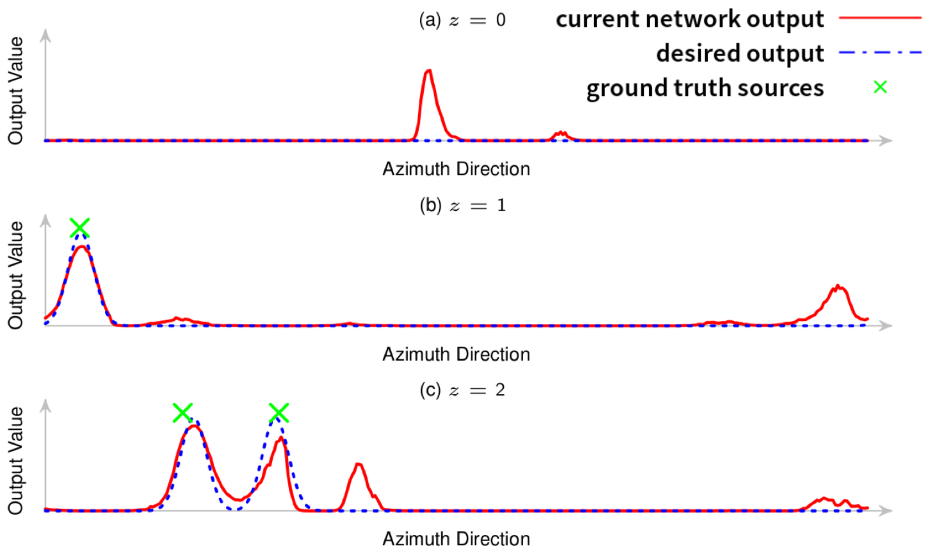
Domain adaptation – weak supervision

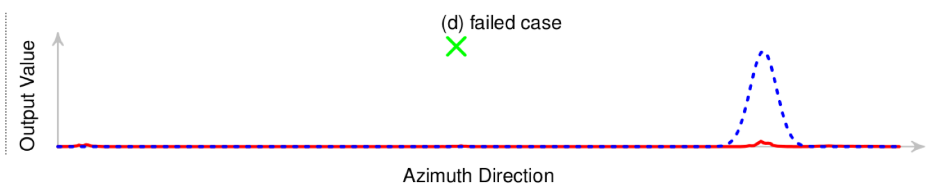
Real sample – Prediction made by DNN trained from synthetic data

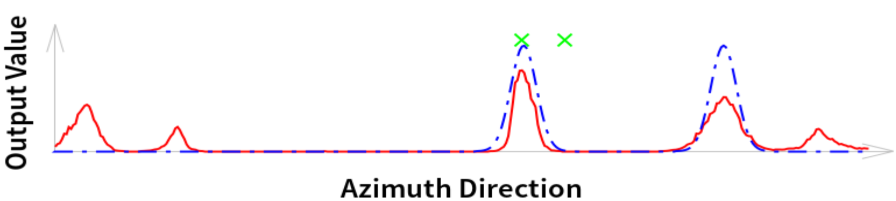




Domain adaptation – weak supervision







- Adaptation
  - real samples
  - generate outputs from the network trained from synthetic data
  - collect desired outputs using the weak label information

$\Longrightarrow$ fine-tune the network with the (real sample, desired output) dataset

Experiments

- Data : clean segment from the AMI corpus
  - Source domain: simulation with RIR generator

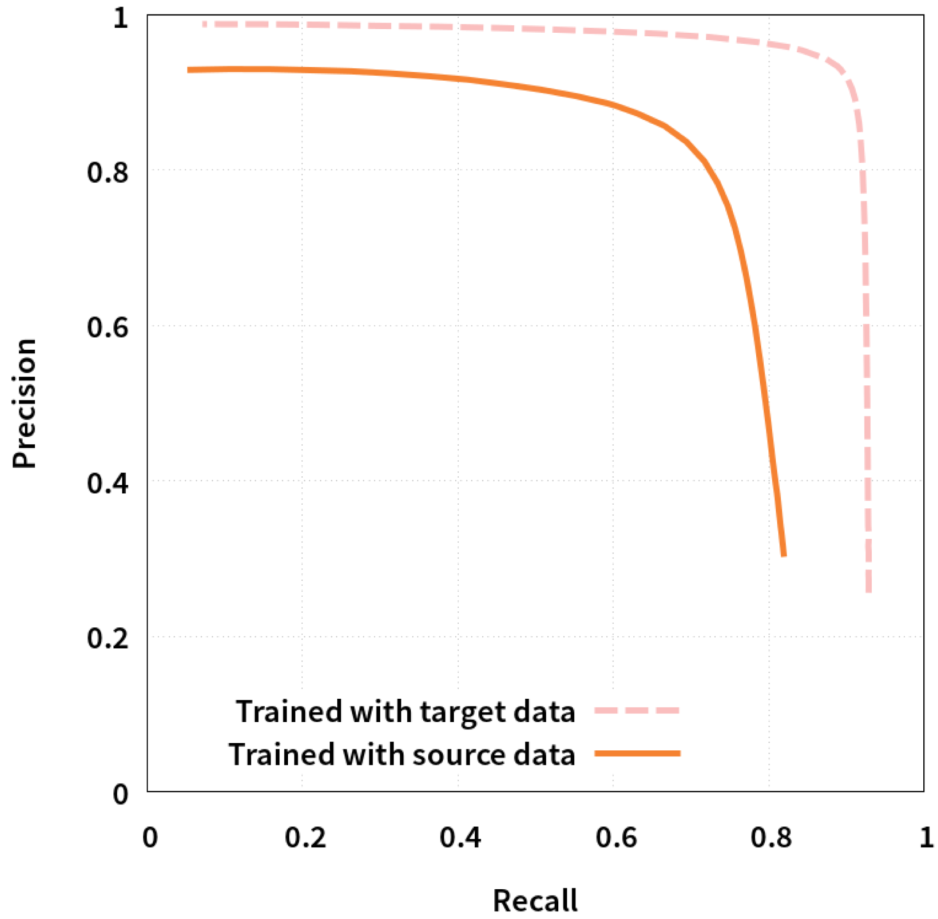
    (several rooms & reverberation coefficients)
  - Target domain: real data (loudspeaker + robot)

- Evaluation
  - Frames of 170ms
  - Maximum 2 sources
  - Correct detection: error < 5 degrees

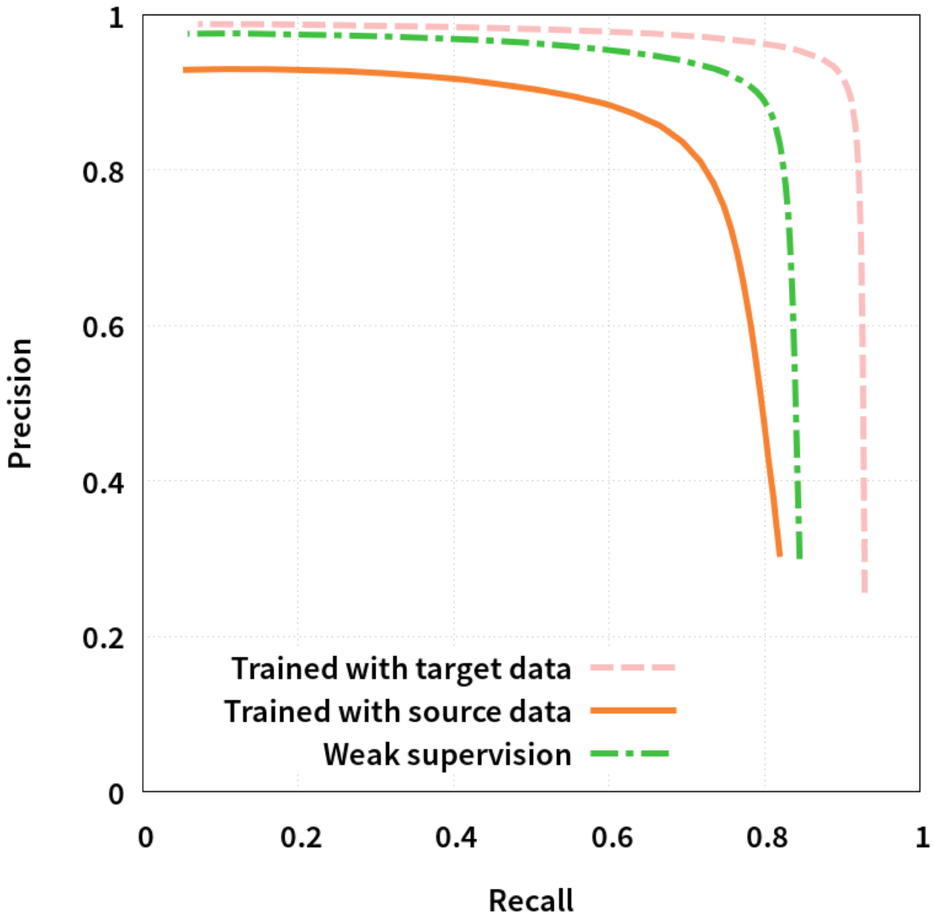
RIR: Room-Impulse-Response simulator .

J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," Journal Acoustic Society of America, 65(4), April 1979, p 943.

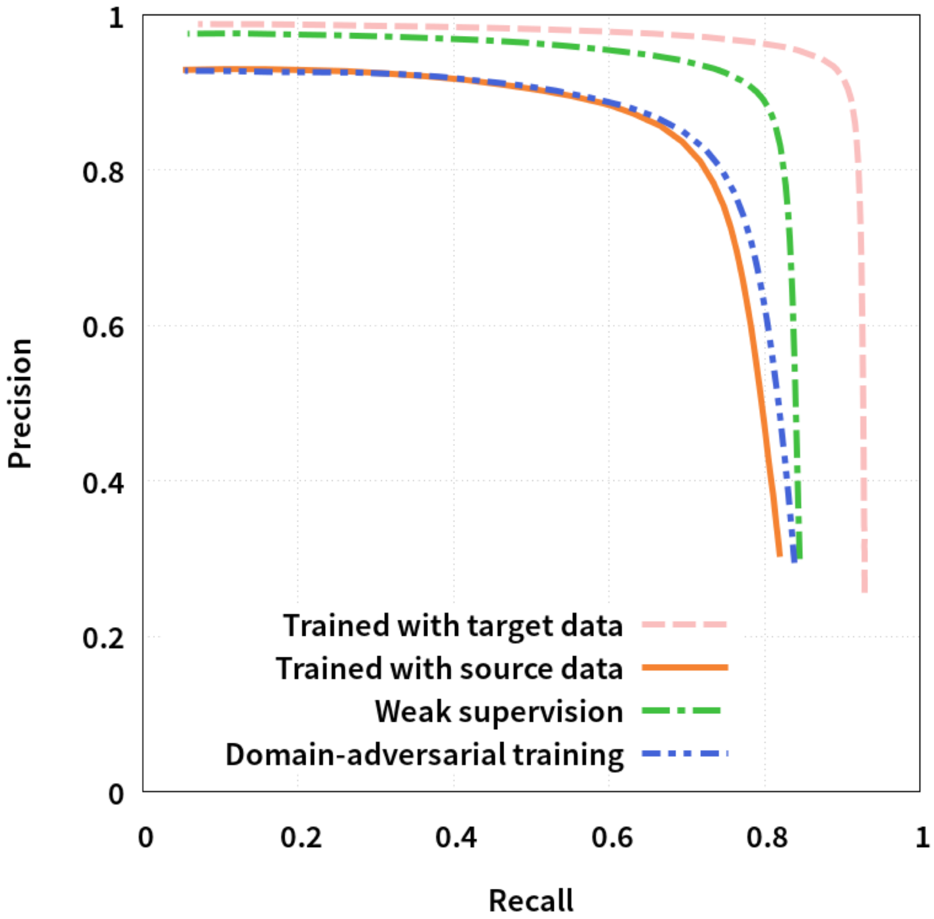
Results – supervised training



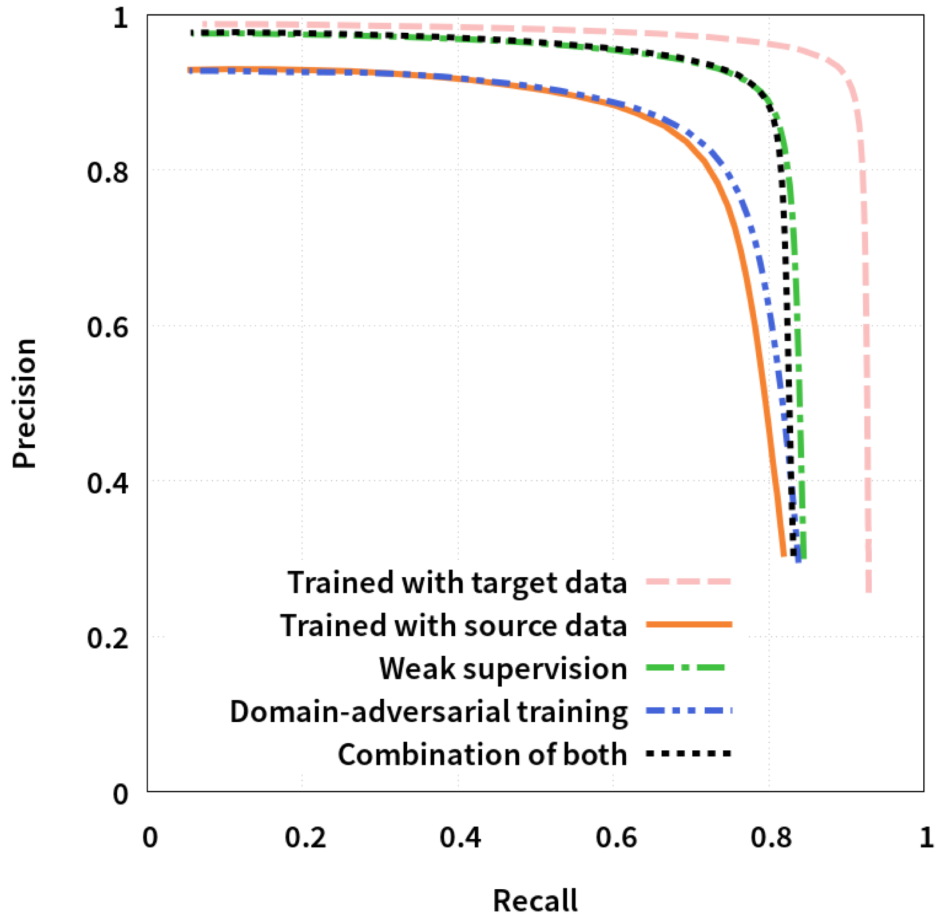
Results – weak supervision



Results - domain adversarial



Results – weak + adversarial combined



- Adversarial training
  - difficult to find the trade-off between domain-invariance and discriminative power

SLS & SNS : conclusions

- Deep learning-based methods
  - Tasks: Sound localization ---- **joint sound localization and classification**
  - **Likelihood** output **encoding** : easy handling of multiple sources
  - **Two-stage training**: adding intermediate supervision
  - **Robot embodiment**: simplifies training data collection
  - Easy addition of temporal context
  - Significant better performance compared to baselines

- Domain adaptation (**quick generalization to other devices**): synthetic to real domain
  - **Weak supervision** (known number of sources) => significant improvement
  - Domain-adversarial training fails to yield significant results

- Next steps
  - Curriculum learning (Done => closes the reality gap)
  - Working **on joint localization & voice embedding**
  - Better simulators

Thanks for your Attention ! Questions ?