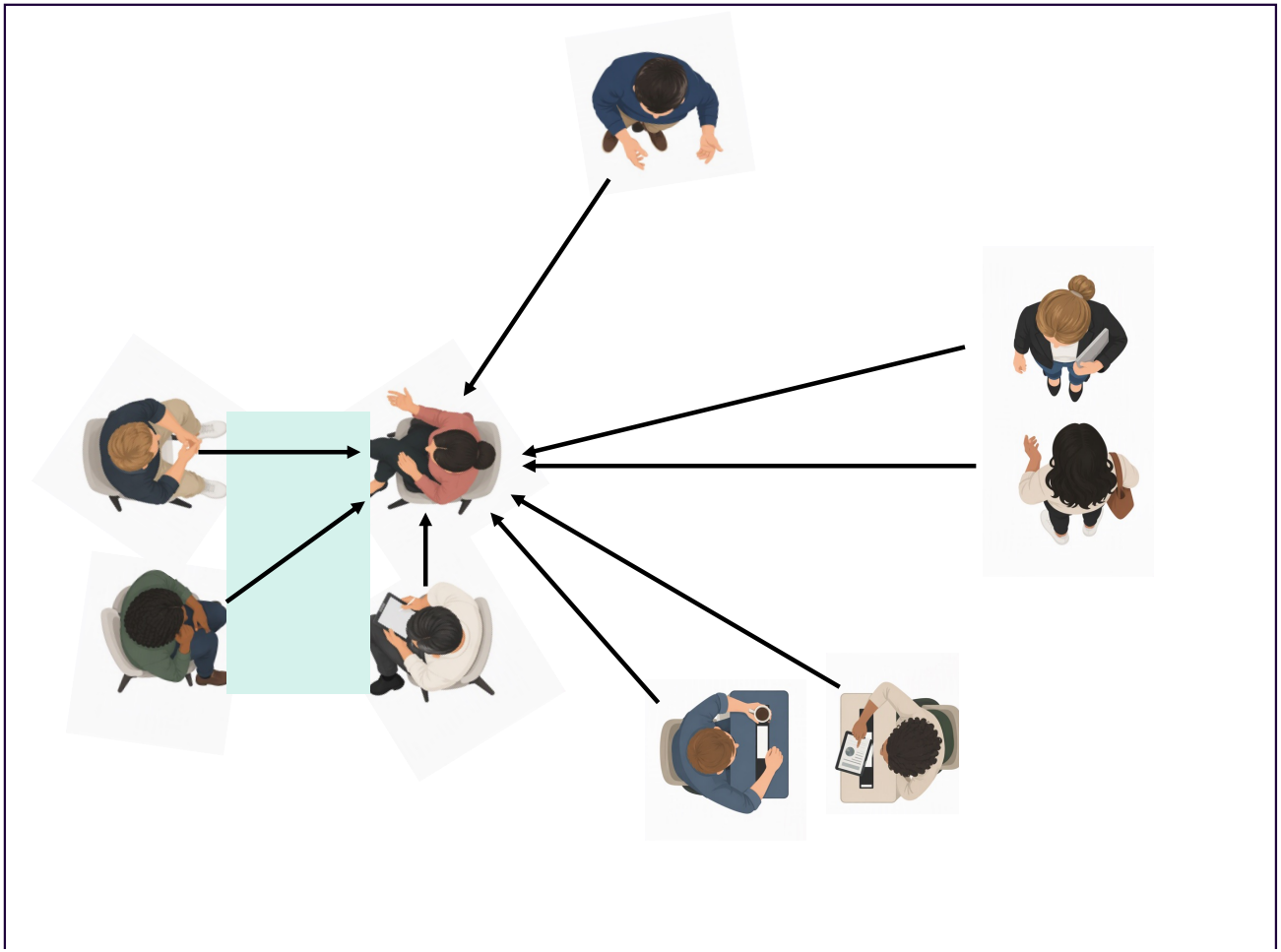
Tampere University

Spatial Analysis of Acoustic Scenes

Tuomas Virtanen
Tampere University

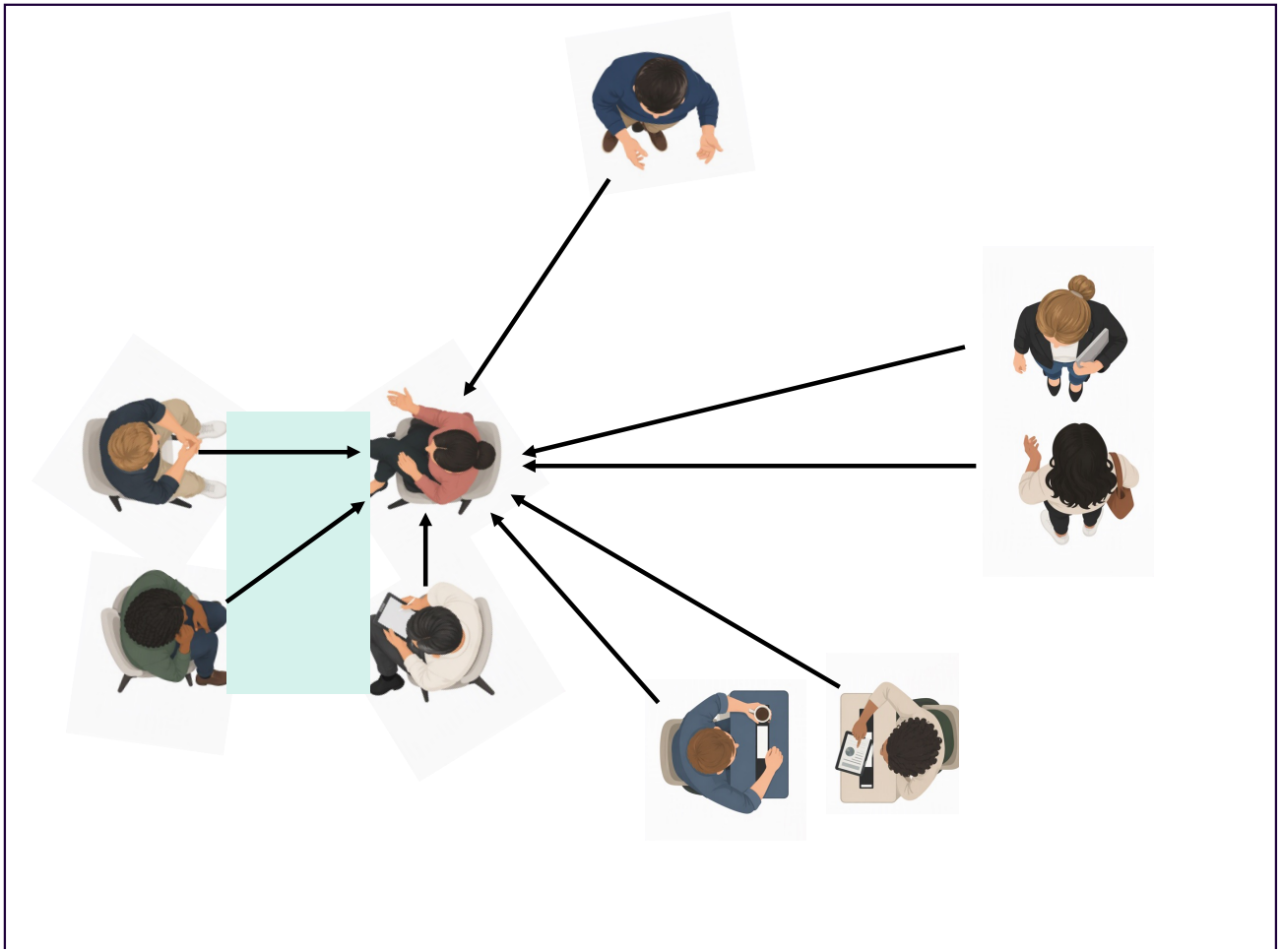
Tampere University

Speech in realistic environments



Tampere University

Speech in realistic environments



Effects:

- Acoustic propagation, noises
- Multiple speakers (possibly in different communication groups)

Opportunities:

- Analysis of spatial information (position, orientations, etc.)

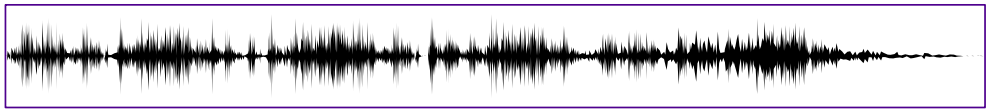
-> can provide cues about attention, communication partnership, and communication structure

- Spatial filtering of audio

Tampere University

Sound event localization and detection

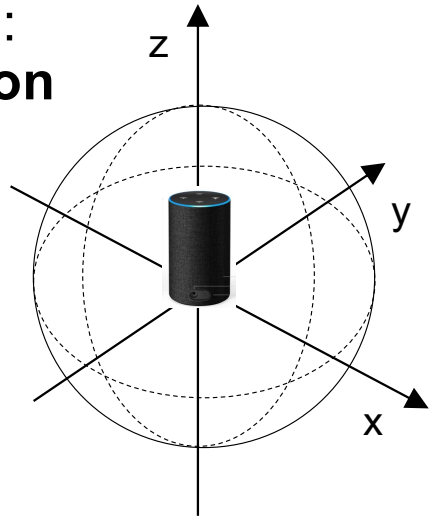






Where are the sound sources:
sound localization
(including distance estimation
and time-varying locations)

What the sources are:
sound event detection

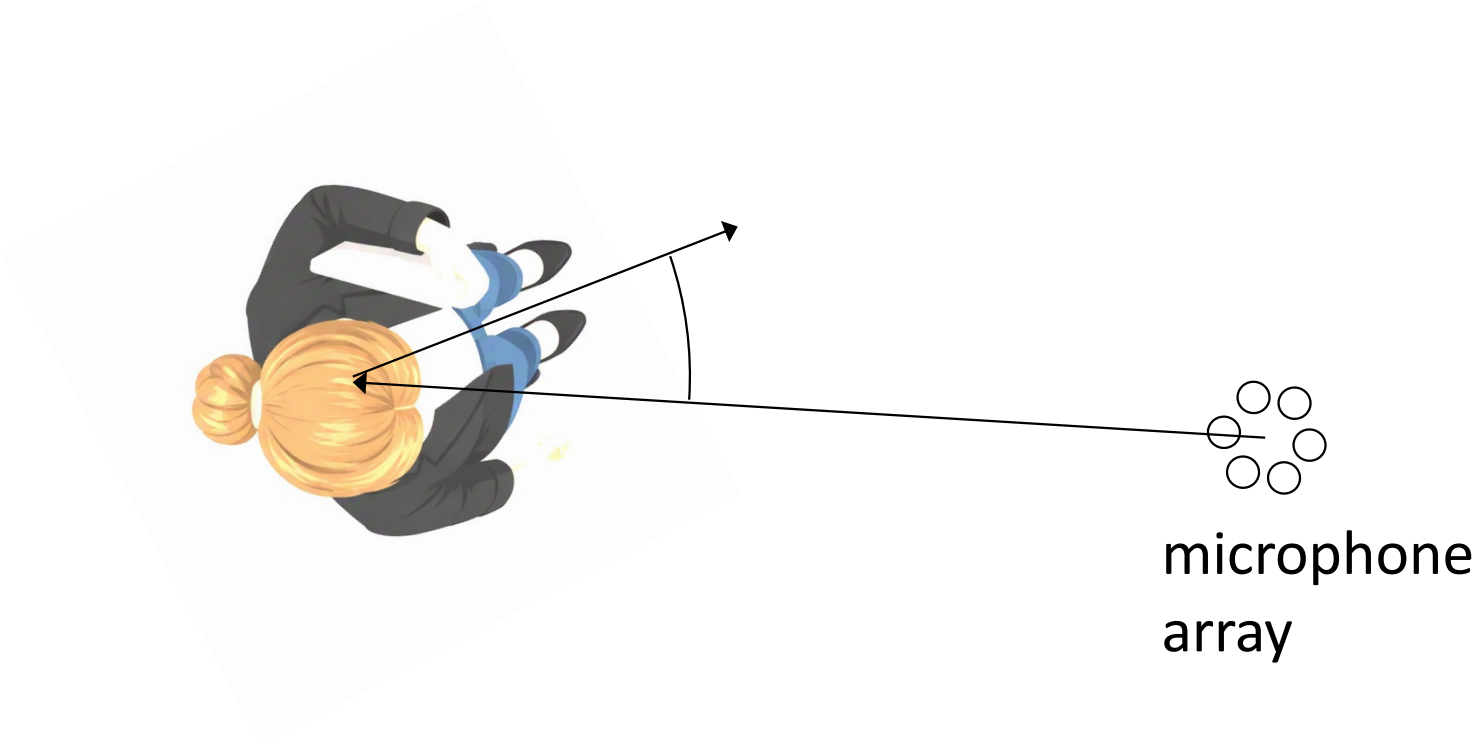


Tampere University



Tampere University

Speaker orientation estimation



Tampere University

Applications

Tampere University

Applications: multimodal assistants & smart environments

- Robots, smart speakers, home appliances, etc.
- Smart environments (education analysis etc.)



Car image by Dllu (CC-SA-BY-4.0)

Tampere University

Applications: wearables / mobile devices

- Hearables (hearing aids, noise-canceling headphones, etc.)
- Mobile phones
- Smart glasses
- Controlling spatial processing, signal enhancement, noise suppression







Hearing aid Image by Michael Thompson (CC-SA-BY 3.0)

Tampere University

Applications: immersive communication

- Controlling audio processing in augmented reality, telepresence, virtual environments..

Bird image by Malene Thyssen, CC BY-SA

Tampere University

Multi-microphone input

- Spatial analysis typically requires a microphone array
- Available in many devices









Multi-microphone input

- Methods developed using supervised learning are specific to certain microphone array
- How to generalize to other devices?
- Device-independent signal representations (Ambisonics)?

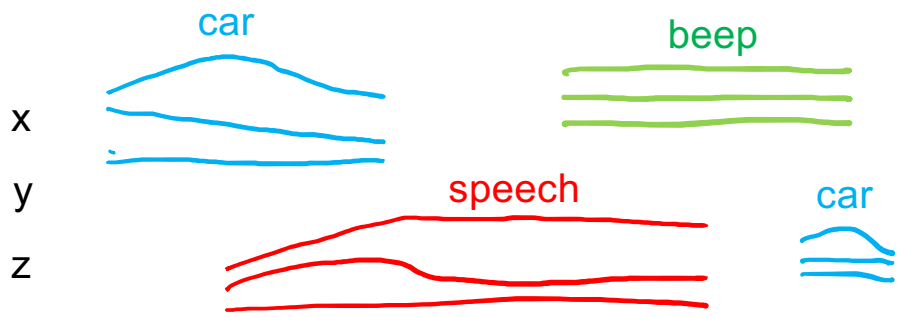
Localization methods: supervised learning

- Training data: input audio + target output
- Set of possible sound classes defined in advance

input audio



target output



Multichannel acoustic features

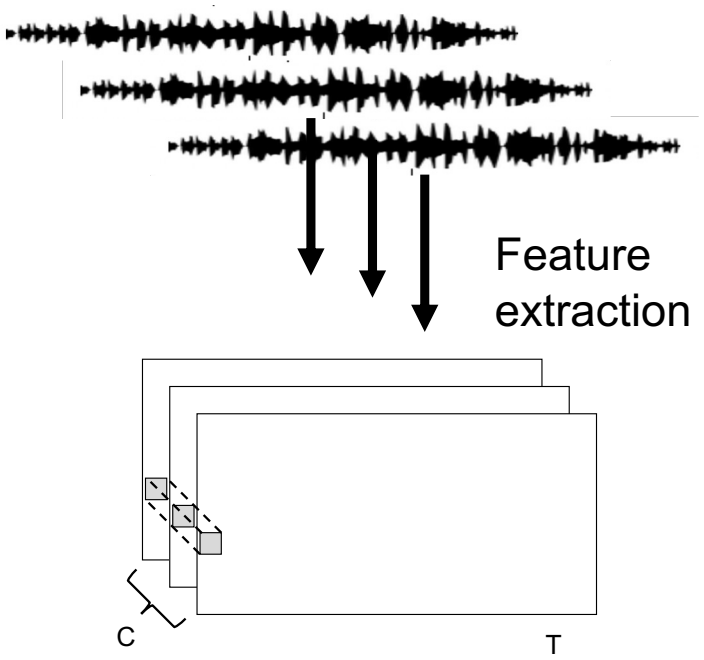
- Explicit feature extraction typically used since end-to-end learning from raw audio computational expensive
- Basic low-level features: STFT magnitude and phase spectrograms (represented using sines & cosines)
- Acoustic intensity vectors, eigenvectors of the spatial covariance matrix
- Generalized cross-correlation between channels

…

Tampere University

Convolutional layers for feature learning

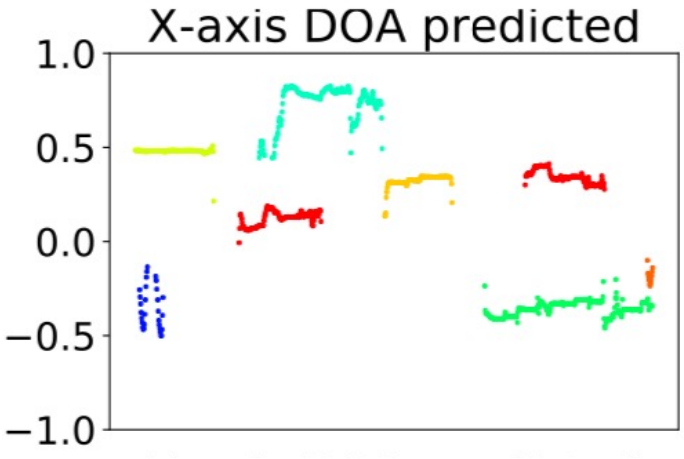
- CNNs for learning features from the low-level feature input
- Convolutional layers enable learning cross-channel information (time differences, level differences) from channel-wise low-level features

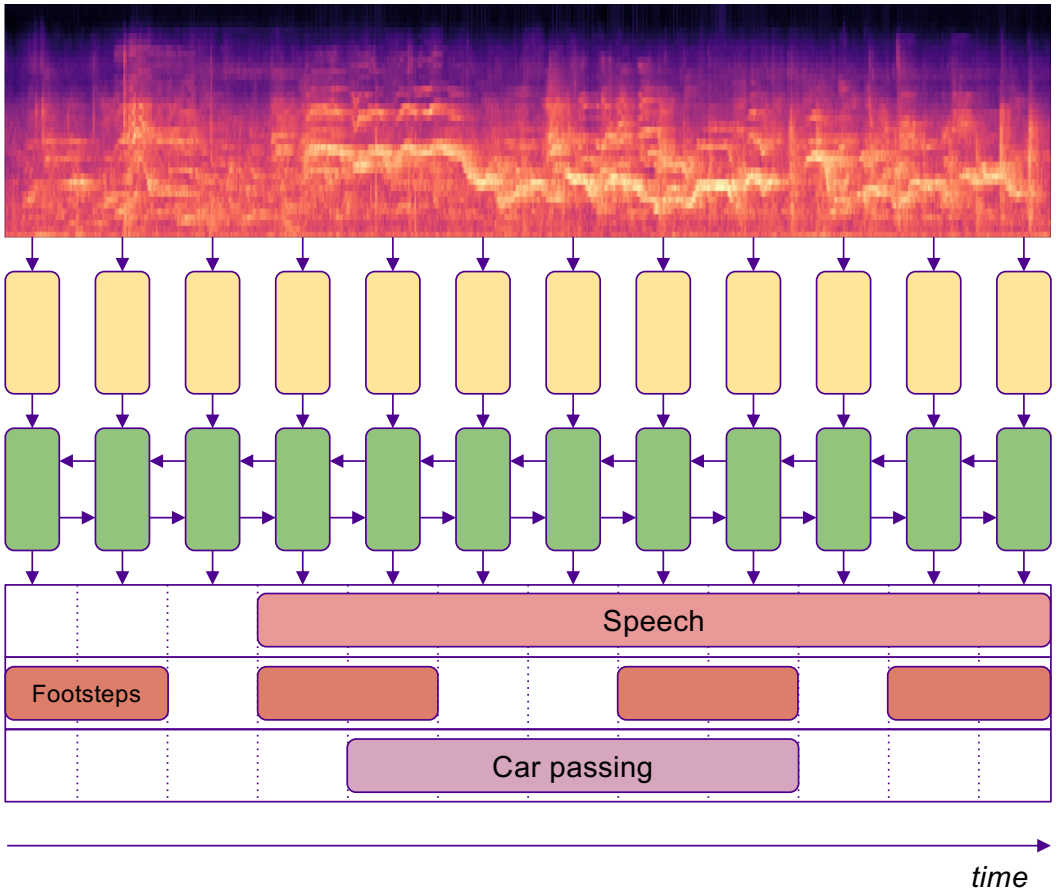


Tampere University

Sequence modeling

- Modeling temporal dynamics important especially in tracking
- Recurrent neural networks, transformers, conformers

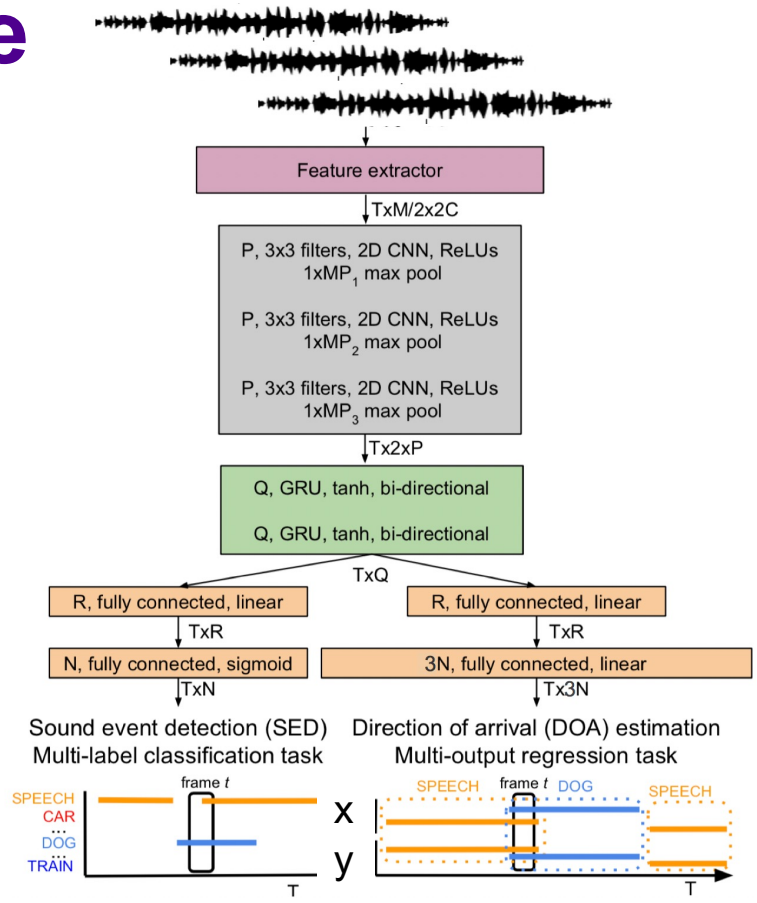




Tampere University

Baseline CRNN architecture

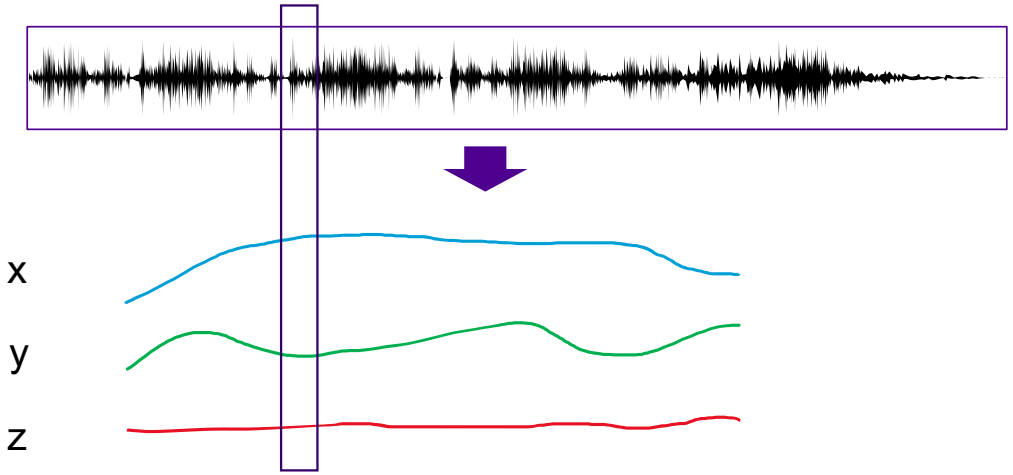
- CNN + RNN layers for feature learning
- Separate fully connected layers for the localization and detection outputs
- The whole network trained jointly in end to end manner
- Masked MSE loss (regression in segments where sources are active)



S. Adavanne, A. Politis, J. Nikunen, T. Virtanen. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. IEEE Journal of Selected Topics in Signal Processing, volume 13, issue 1, 2019.

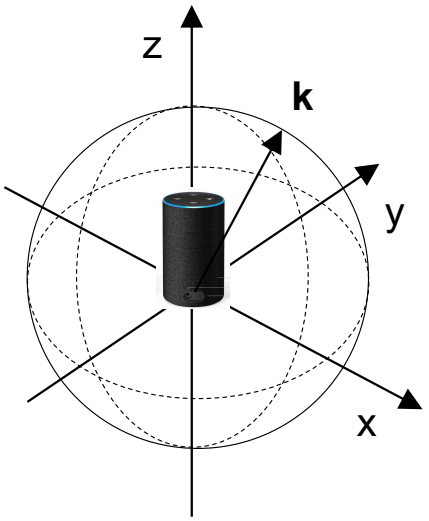
Neural network output

- Single-source localization: coordinates (x, y, z) per each time frame



Output coordinate space

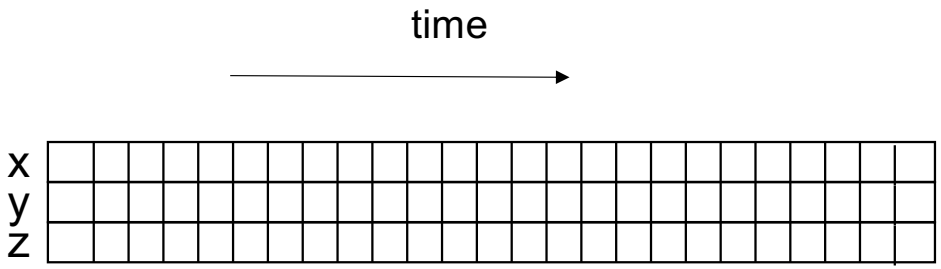
1. Direction of source: unit norm vector **k** = $[x,y,z]^T$, or horizontal and vertical angle
2. Position in cartesian coordinates $[x,y,z]^T$
3. Direction + distance



Tampere University

Tracking

- Sources may not be temporally persistent (e.g. there are pauses in speech)
- No sound → nothing to locate: what will be the regression output?

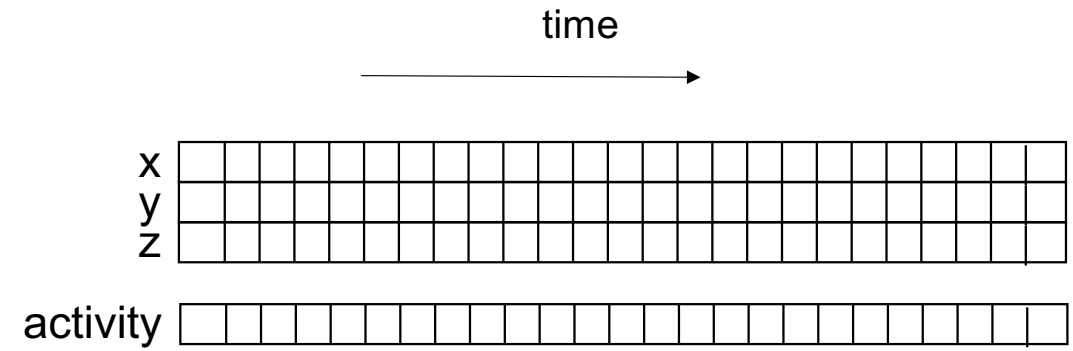


Tampere University

Tracking

- Sources may not be temporally persistent (e.g. there are pauses in speech)
- No sound → nothing to locate: what will be the regression output?

1. Separate output to indicate the activity



Tampere University

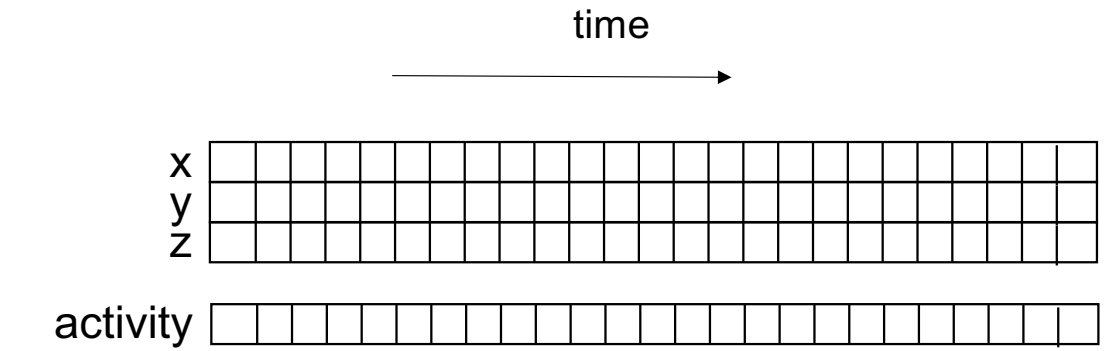
Tracking

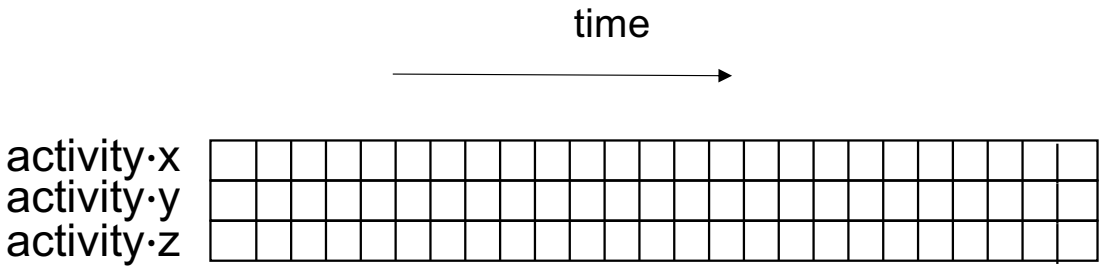
- Sources may not be temporally persistent (e.g. there is pauses in speech)
- No sound → nothing to locate: what will be the regression output?

1. Separate output to indicate the activity

2. Encoding the activity of a source to the magnitudes of the $[x\ y\ z]^T$ direction vector (ACCDOA, Shimada et al. 2020)

- Loss computed only for active frames

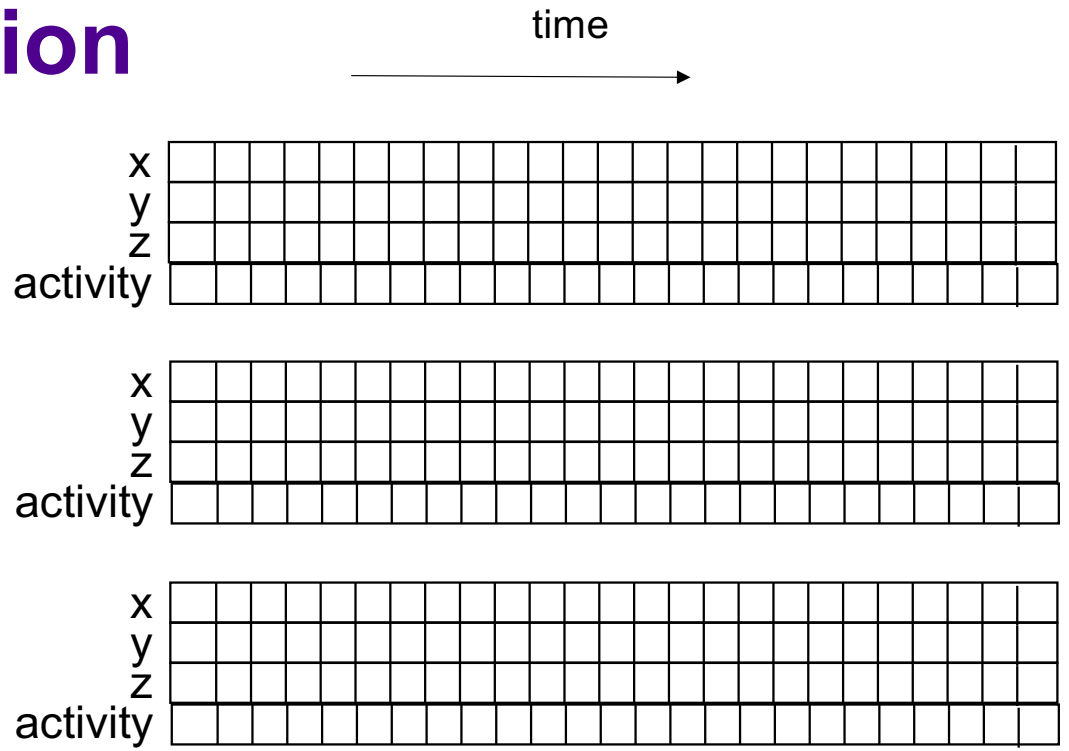




Tampere University

Multi-source localization

- x_i, y_i, z_i for each source i to be localized



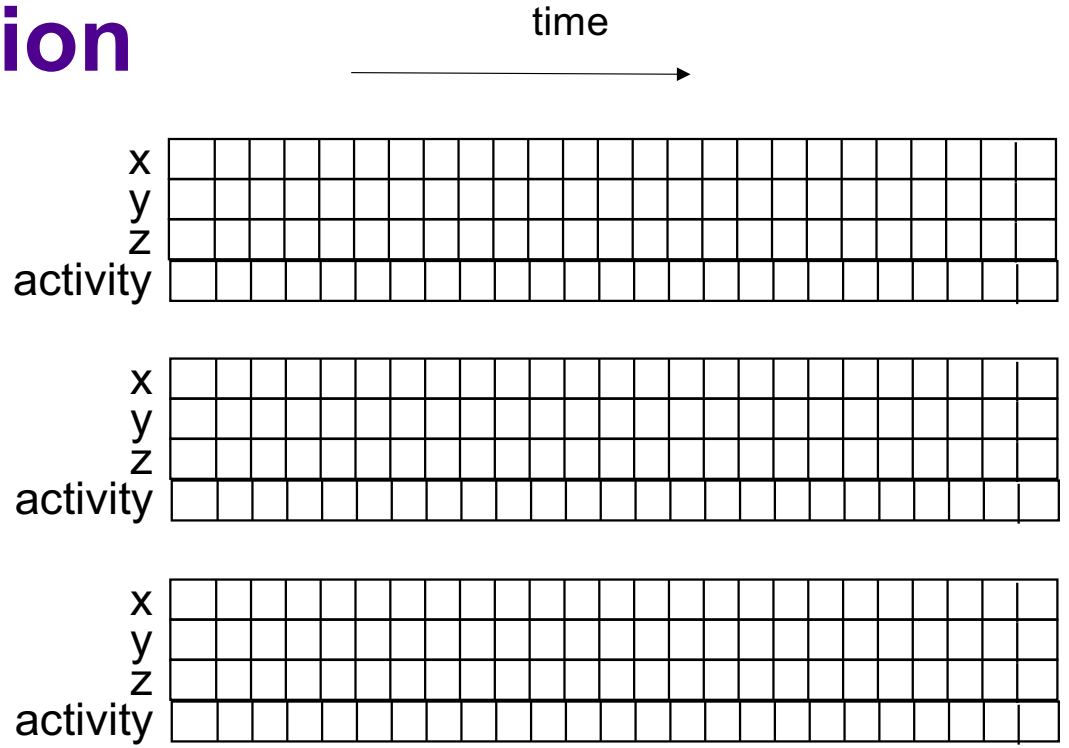
Tampere University

Multi-source localization

- x_i, y_i, z_i for each source i to be localized
- Need to handle the permutation of sources in training and evaluation

➔permutation-invariant training (PIT)

- Try all combinations of predictions and references, select the permutation that has lowest loss

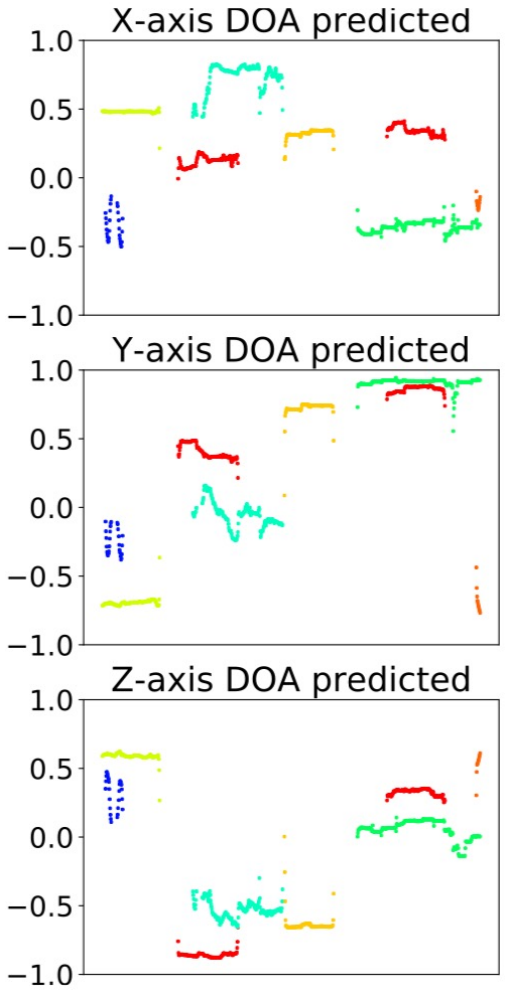


S. Adavanne, A. Politis, and T. Virtanen. Differentiable Tracking-Based Training of Deep-Learning Sound Source Localizers. In proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2021.

Tampere University

Outputting estimated source locations

- Masking x, y, z coordinates by the activities



Tampere University

Data

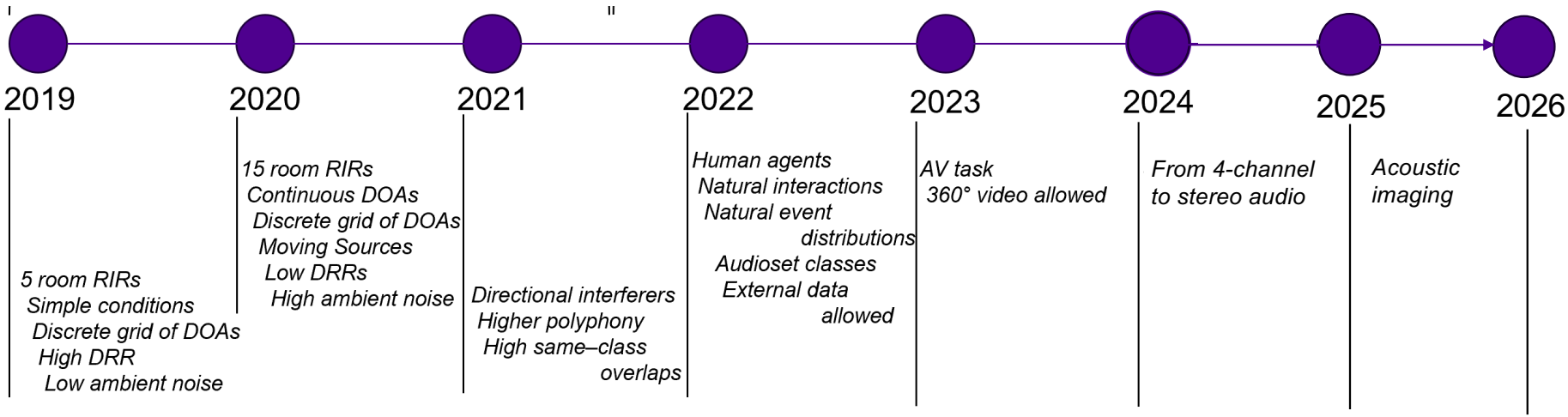
- The main factor affecting the performance of the developed systems
- Several factors affecting the captured audio:
  - Geometry of the microphone array
  - Microphone responses
  - Transfer function from source to microphone
  - Interfering sources
  - Properties sources themselves
- How to deal with all the variability?

Obtaining data

1. Real recordings
   - Audio is relatively easy to record, annotations cumbersome
   - Match with real scenarios
2. Simulating material
   - Mixing of target class samples and background
   - Simulating impulse responses corresponding to source locations
   - Easy to produce lots of data and obtain annotations
   - Do the results translate to real environments?
3. Hybrid approaches
   - E.g. real impulse responses + background noises

Tampere University

Example: DCASE 2019-2026



Tampere University

Capture setup

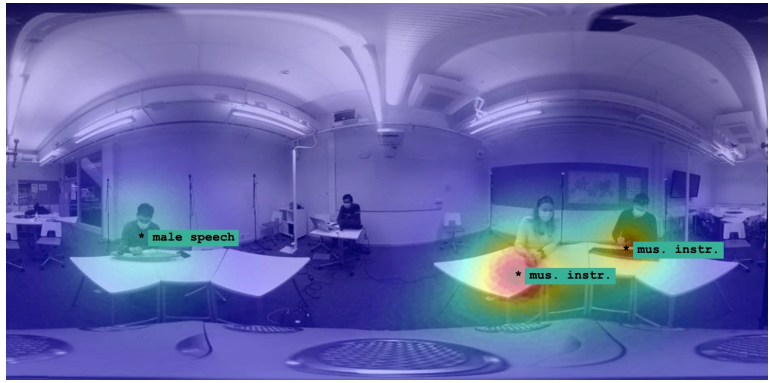
- Eigenmike (32 audio channels)
- Ricoh theta 360 degrees video (for reference)



Tampere University

STARSS23 dataset

- Audio-visual data with spatiotemporal annotations of sound events from 13 classes
- About 11hrs of scenes recorded in Tampere (TAU) and Tokyo (SONY)
- 4-microphones, FOA (captured with 32-channel Eigenmike)



K. Shimada, A. Politis, P. Sudarsanam, D. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y Mitsufuji. STARSS23: An Audio-Visual Dataset of Spatial Recordings of Real Scenes with Spatiotemporal Annotations of Sound Events. In proc. NeurIPS 2023.

Tampere University

Data collection

- Real scenes with recruited actors recorded in diverse indoor environments
- Automatic 3D localization of the sound sources using Optitrack motion capture system
- Close-miked sources to allow (semi-)automatic activity estimation
- 360 video recorded as a reference

Tampere University

Hardware



RICOH THETA V
(360 video)



Em32 Eigenmike
(Spatial audio)



Optitrack Motion Capture
system (3D-localization)



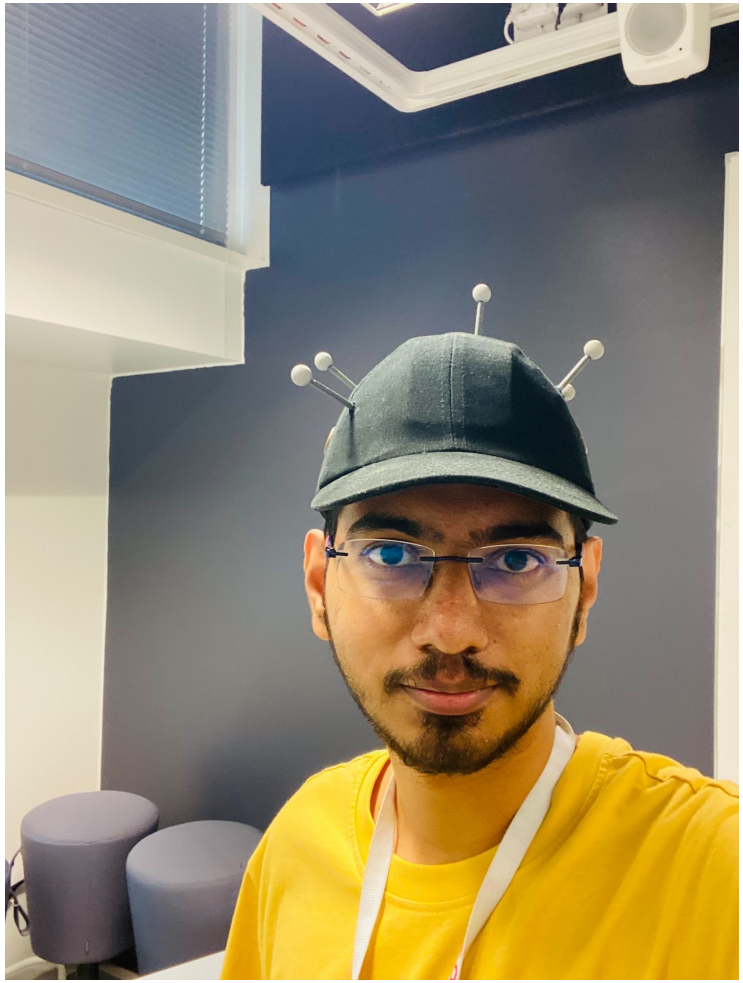
RODE Wireless Go II
(close miked sources)

Tampere University

Data collection set up



Tampere University





Tampere University

Locations



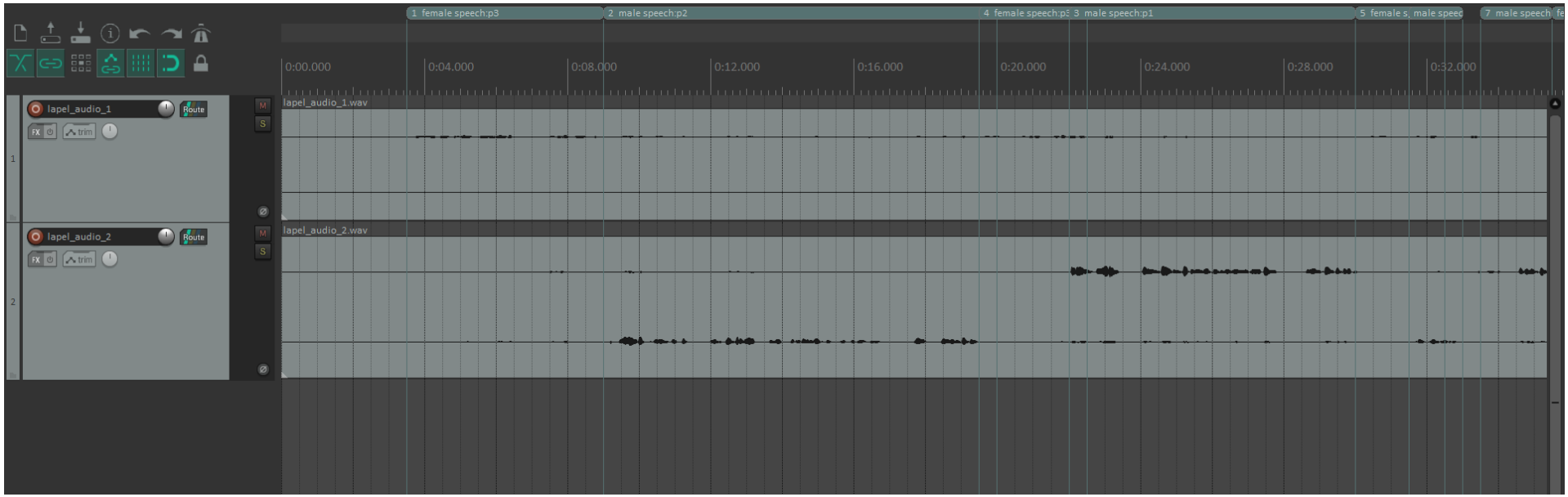




Tampere University

Sound event annotation

- Automatic annotation of activities based on close-miked signals
- Manual checking and correction




malSpeech

Tampere University

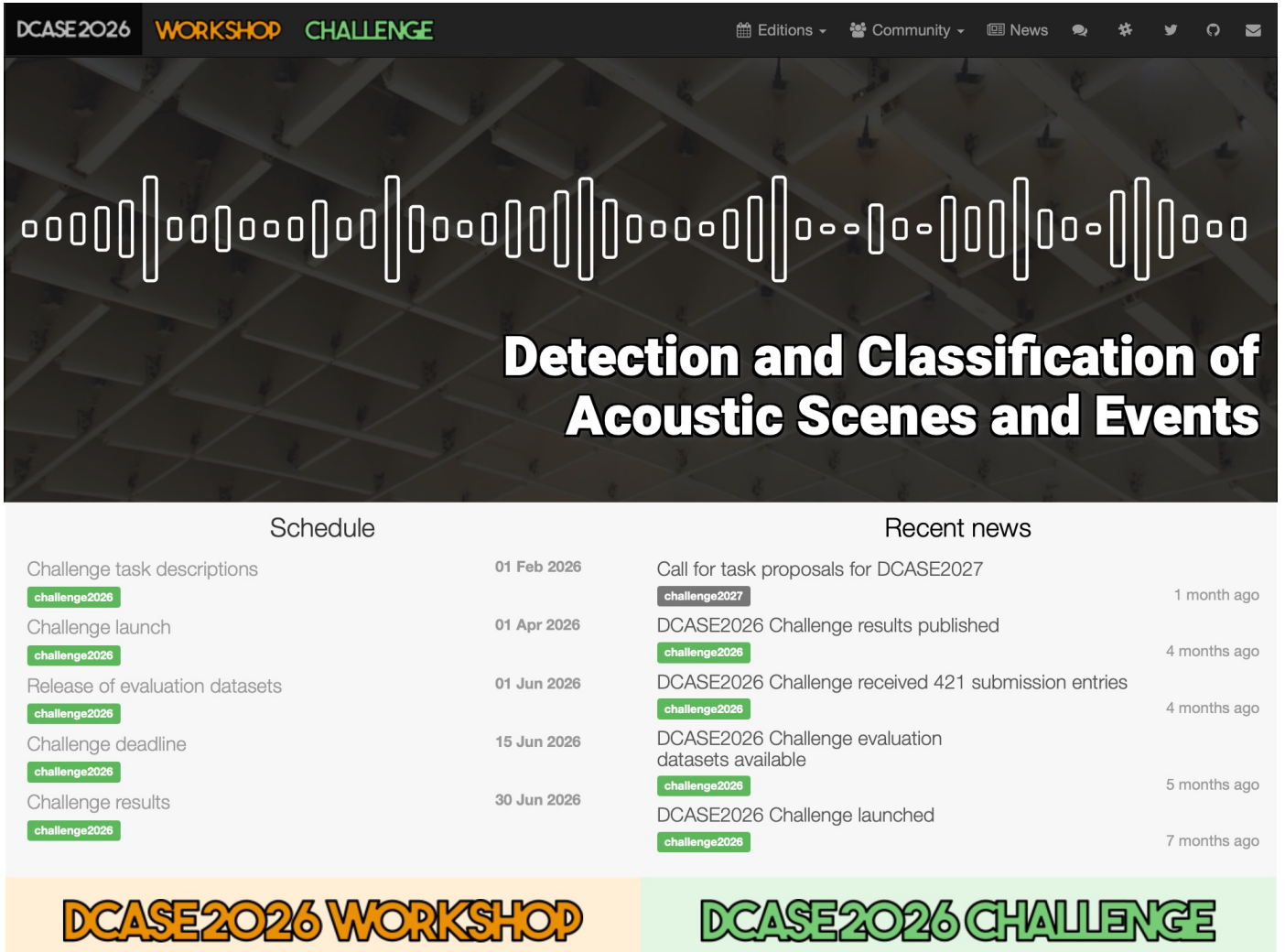
Audio classes & coverage

	Global	Fem. speech	Male speech	Clap	Phone	Laugh	Dom. sounds	Footsteps	Door	Music	Music. instr.	Faucet	Bell	Knock
Frame coverage (% total frames)	82.5	28.4	31.4	0.53	1.02	2.89	18.6	2.17	0.70	22.2	2.32	0.58	0.99	0.06
Max. polyphony	6	2	3	2	1	4	2	3	1	2	2	1	1	1
Mean polyphony	1.47	1.05	1.06	1.06	1.00	1.25	1.03	1.02	1.00	1.20	1.28	1.00	1.00	1.00
Polyphony 1 (% active frames)	62.878	95.44	94.35	93.74	100	78.67	97.27	97.87	100	79.99	71.66	100	100	100
Polyphony 2	28.443	4.56	5.51	6.26	0	18.48	2.73	1.94	0	20.01	28.36	0	0	0
Polyphony 3	7.385	0	0.14	0	0	2.39	0	0.19	0	0	0	0	0	0
Polyphony 4	1.129	0	0	0	0	0.46	0	0	0	0	0	0	0	0
Polyphony 5	0.162	0	0	0	0	0	0	0	0	0	0	0	0	0
Polyphony 6	0.003	0	0	0	0	0	0	0	0	0	0	0	0	0

Tampere University

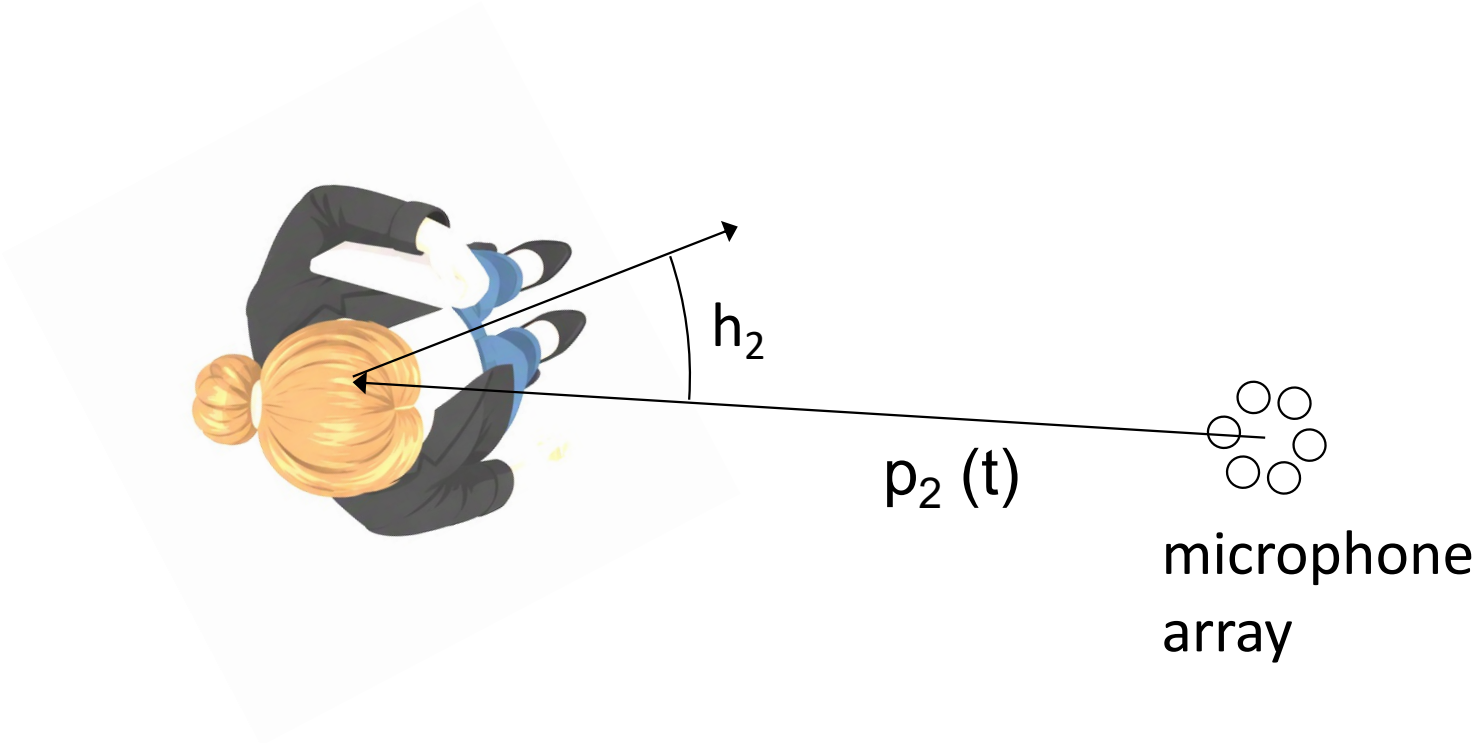
Results

http://dcase.community



Tampere University

Speaker orientation estimation

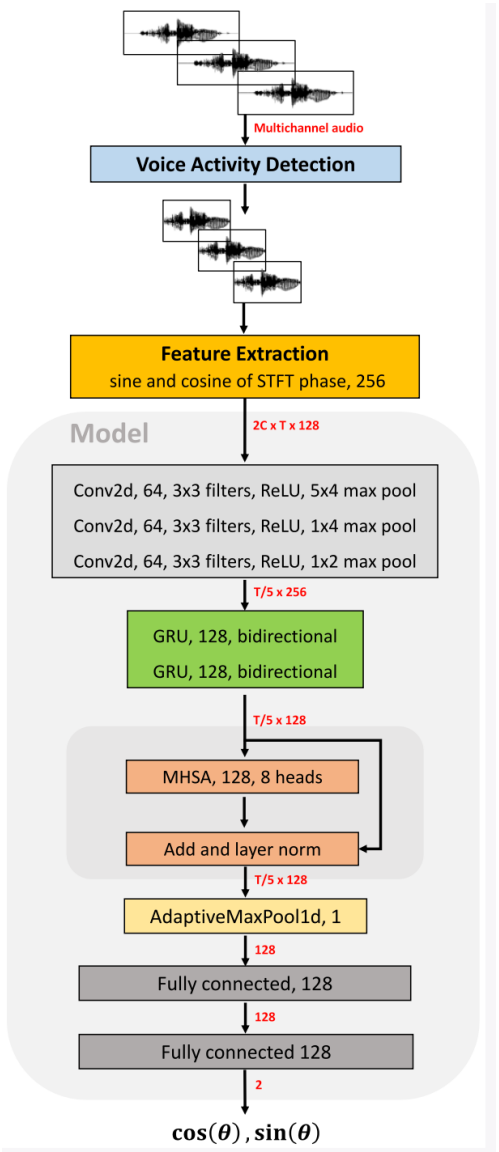


B. Turi, A. Politis, P. Sudarsanam, and T. Virtanen. Speaker Head Orientation Estimation with a Single Microphone Array Using Phase Spectrogram Features. In proc. European Signal Processing Conference 2026.

Tampere University

Orientation estimation network

- A convolutional-recurrent-attention network maps multichannel phase features to a circular orientation representation

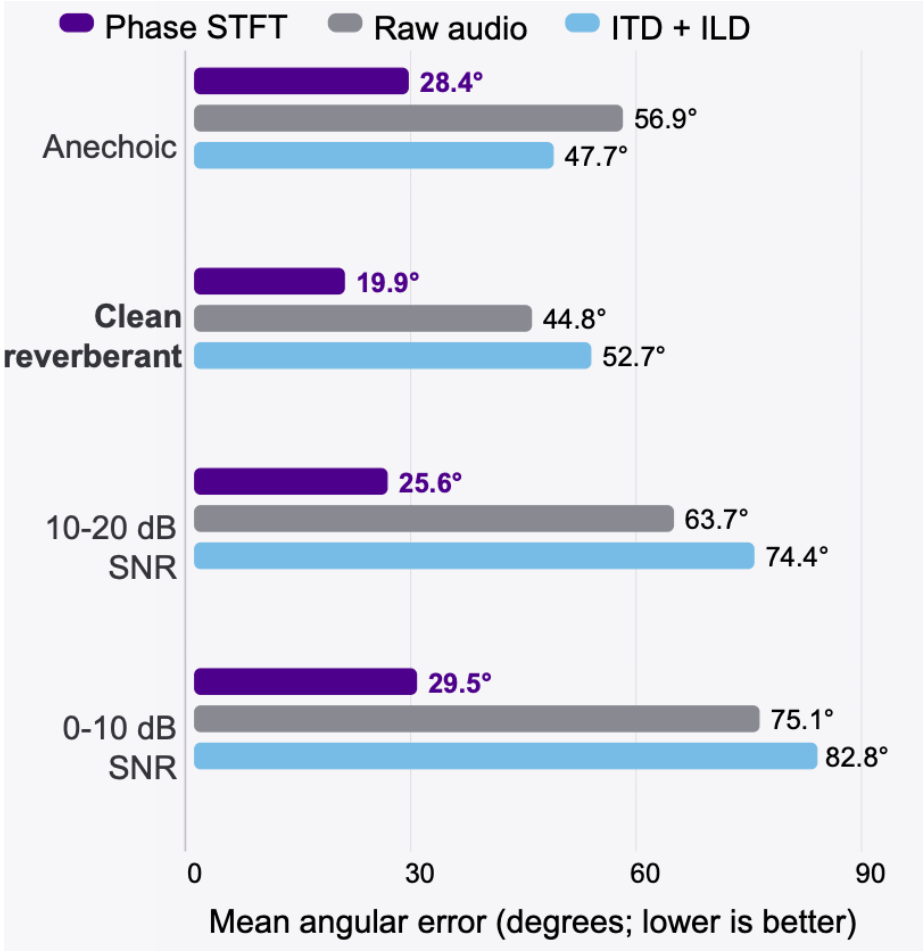


B. Turi, A. Politis, P. Sudarsanam, and T. Virtanen. Speaker Head Orientation Estimation with a Single Microphone Array Using Phase Spectrogram Features. In proc. European Signal Processing Conference 2026.

Tampere University

Orientation results

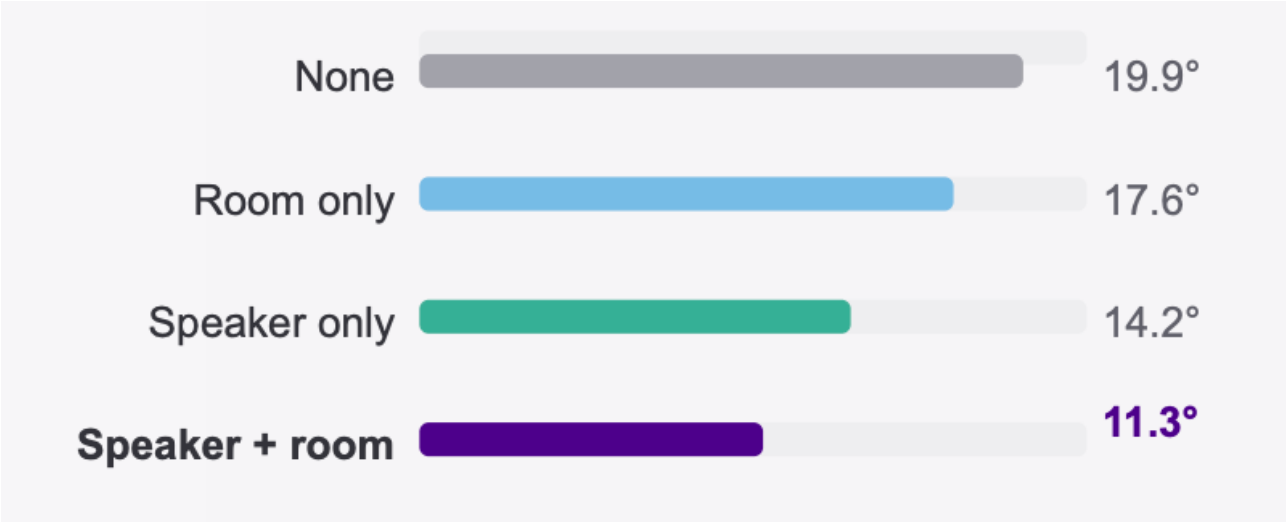
- Reverberation helps orientation estimation in comparison to anechoic signals!



Tampere University

Personalization

- Fine-tuning the model to specific room or for specific speaker



Tampere University

Discussion

Tampere University

Challenges in supervised learning based spatial analysis

- Realistic training and evaluation data bottleneck in supervised learning
  -> still need for ways to collect and use data more efficiently
- Producing ground-truth information about locations very difficult (manual annotation or the use of other sensors)

Tampere University

Challenges in supervised learning based spatial analysis

- Training material is captured with a specific microphone array -> trained methods are also device-specific
- How to adapt models to other devices?

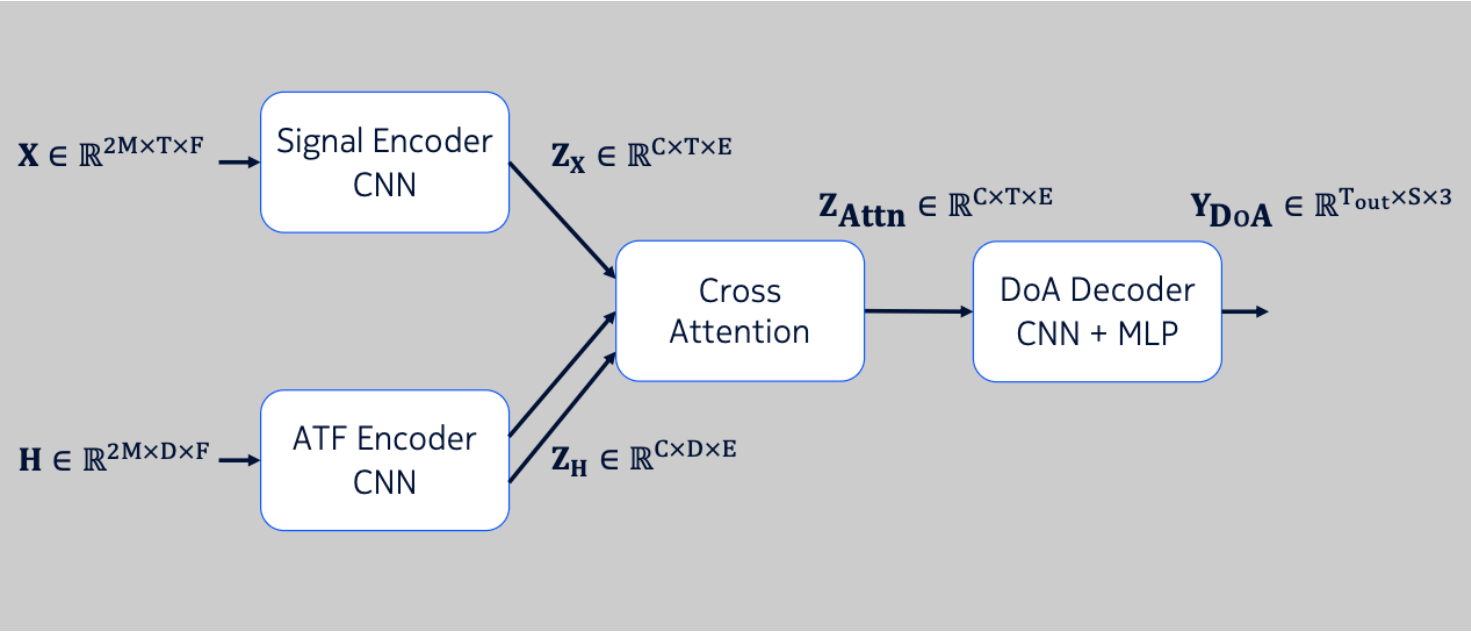
Microphone array generalization

- Captured spatial cues depend on the geometry of the microphone array:
- Number of microphones, their positions
- Also: responses of microphones

Tampere University

Microphone array generalization

- Idea: condition the model by measured microphone responses
- Cross-attention between audio and microphone responses



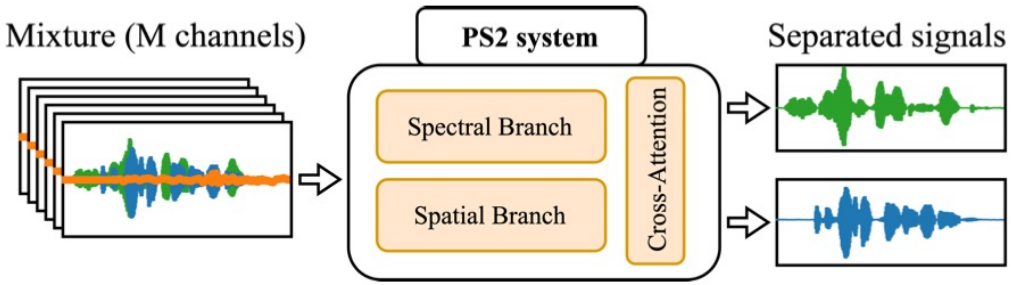
M. Heikkinen, A. Politis, K. Drossos, and T. Virtanen. Neural Array-Generic Direction-of-Arrival Estimation Exploiting Array Transfer Functions. In proc. International Workshop on Acoustic Signal Enhancement 2026.

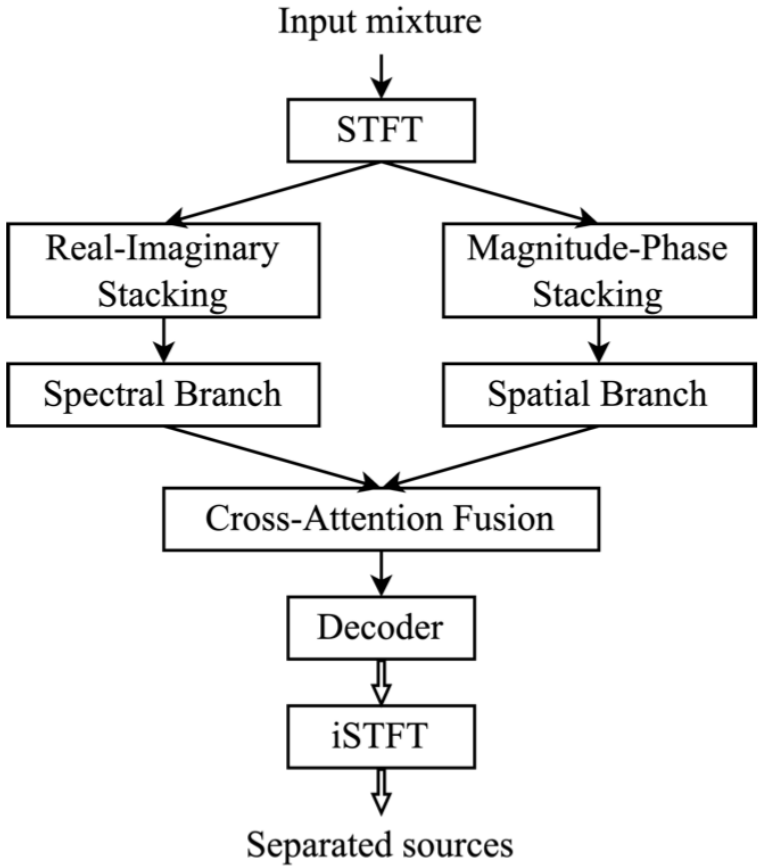
Speaker separation

- Spatial filtering vs. learned models

Speaker separation

- Spatial filtering vs. learned models



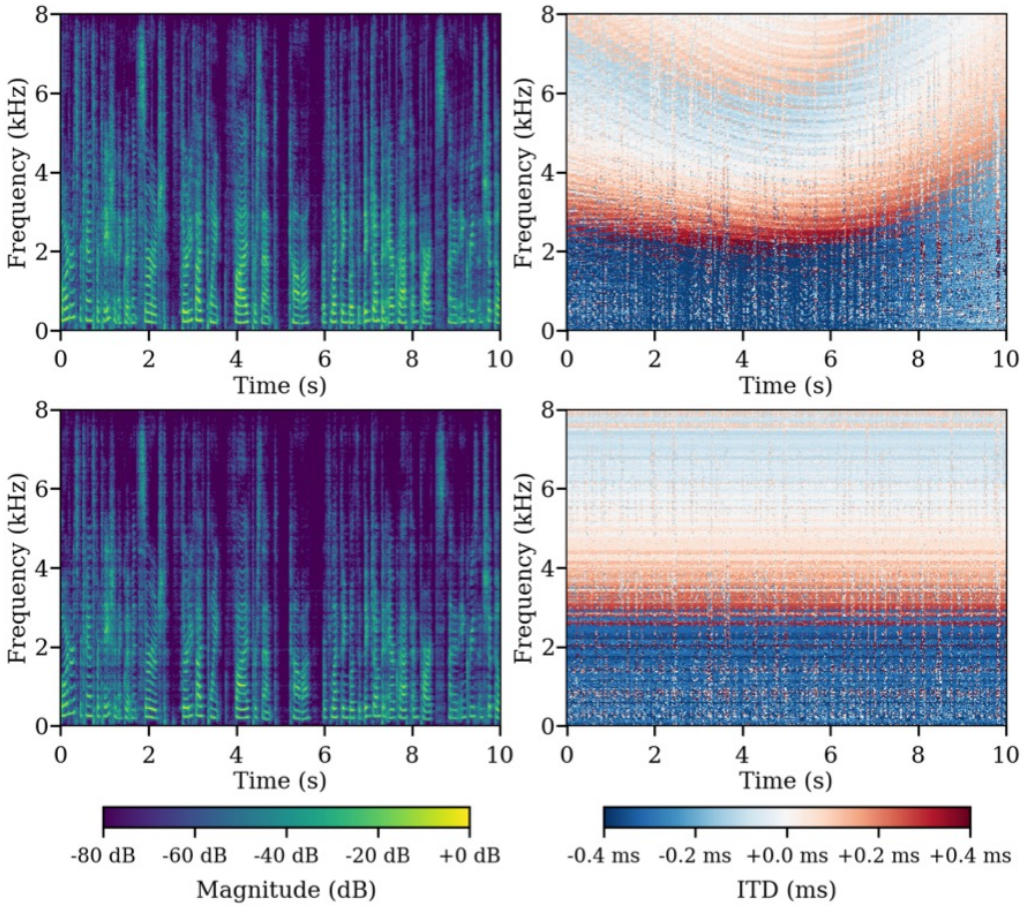


Y. Wang, A. Politis, K. Drossos, and Tuomas Virtanen. Moving Speaker Separation via Parallel Spectral-Spatial Processing. IEEE Transactions on Audio, Speech and Language Processing, Volume 34, 2026.

Tampere University

Speaker separation

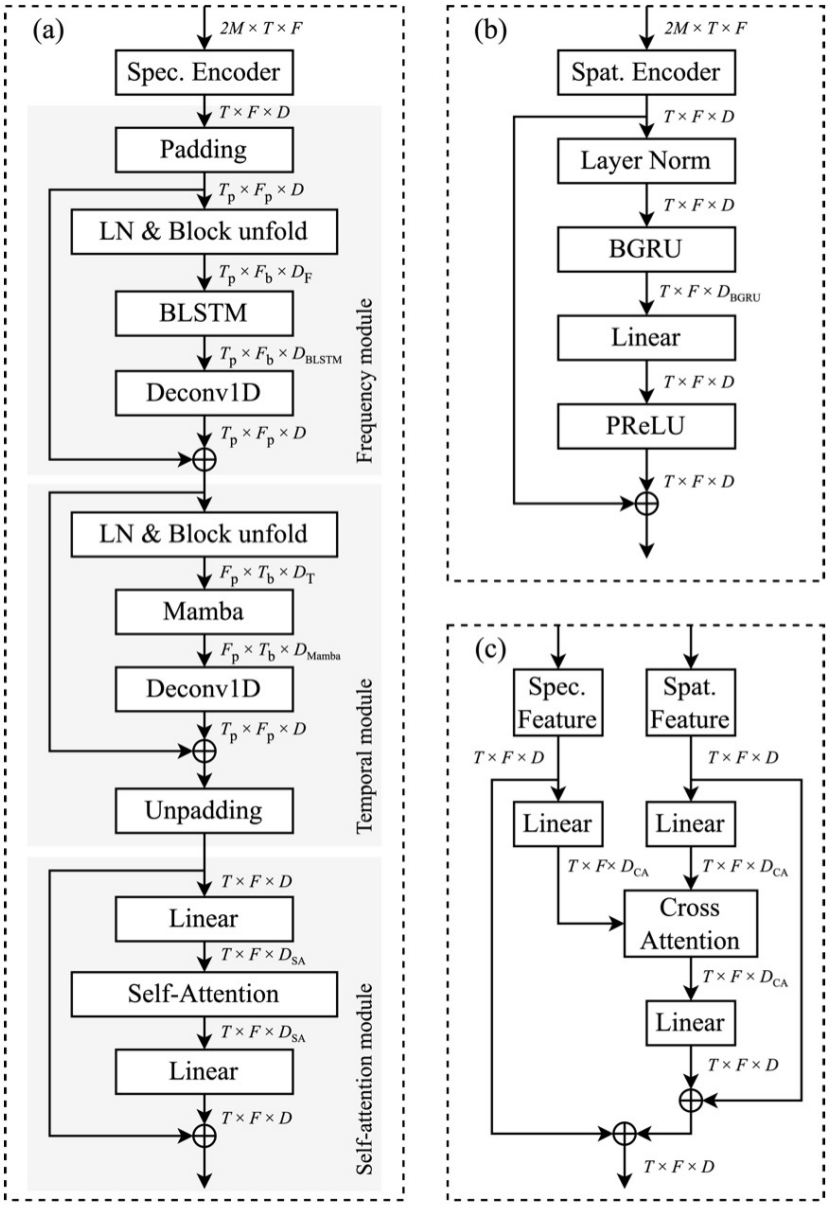
- Spatial filtering vs. learned models



Tampere University

Speaker separation

- Spatial filtering vs. learned models



Tampere University

Opportunities

- Distributed microphones & microphone arrays
- Integration to multimodal language models

Tampere University

Scientific channels

DCASE WORKSHOP

Workshop on Detection and Classification of Acoustic Scenes and Events:

- Annual scientific event covering tasks, methods, resources, applications, and evaluation

http://dcase.community

DCASE CHALLENGE

Annual data challenge related to computational acoustic scene analysis

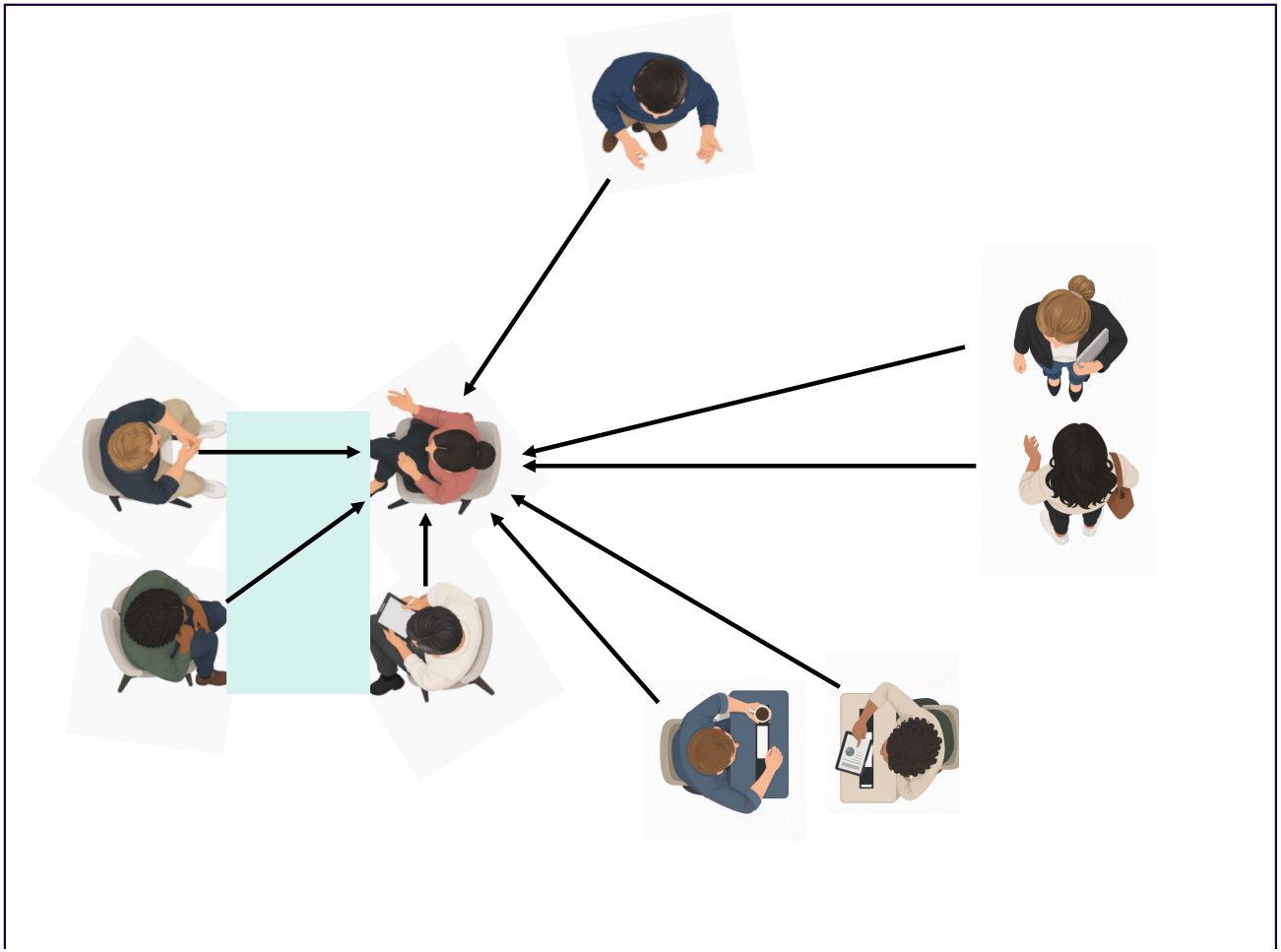
Tampere University

Contributors

- Researchers at ARG/TAU + Sony + Nokia

Tampere University

Summary: speech in realistic environments



Effects:

- Acoustic propagation, noises
- Multiple speakers (possibly in different communication groups)

Opportunities:

- Analysis of spatial information (position, orientations, etc.)

-> can provide cues about attention, communication partnership, and communication structure

- Spatial filtering of audio

Tampere University

Thank you!