ROMA TRE
UNIVERSITÀ DEGLI STUDI

Scene Understanding with Sound using Artificial Intelligence Techniques

Supervisors
Prof. Marco Carli
Prof. Alessandro Neri

Ph.D. Candidate
Michael Neri

XXXVII° cycle - Ph.D. course in Applied Electronics
30th April, 2025

Motivation

Scene understanding is the process **of perceiving, analyzing, and processing data of a 3D dynamic scene** from **a network of sensors.**

Why exploiting **sound** in scene understanding?

Easy to collect and process

More robust to occlusions with respect to vision

Real-time monitoring

Motivation

Artificial intelligence techniques (mostly very complex deep neural networks) are used for **sound scene understanding**.

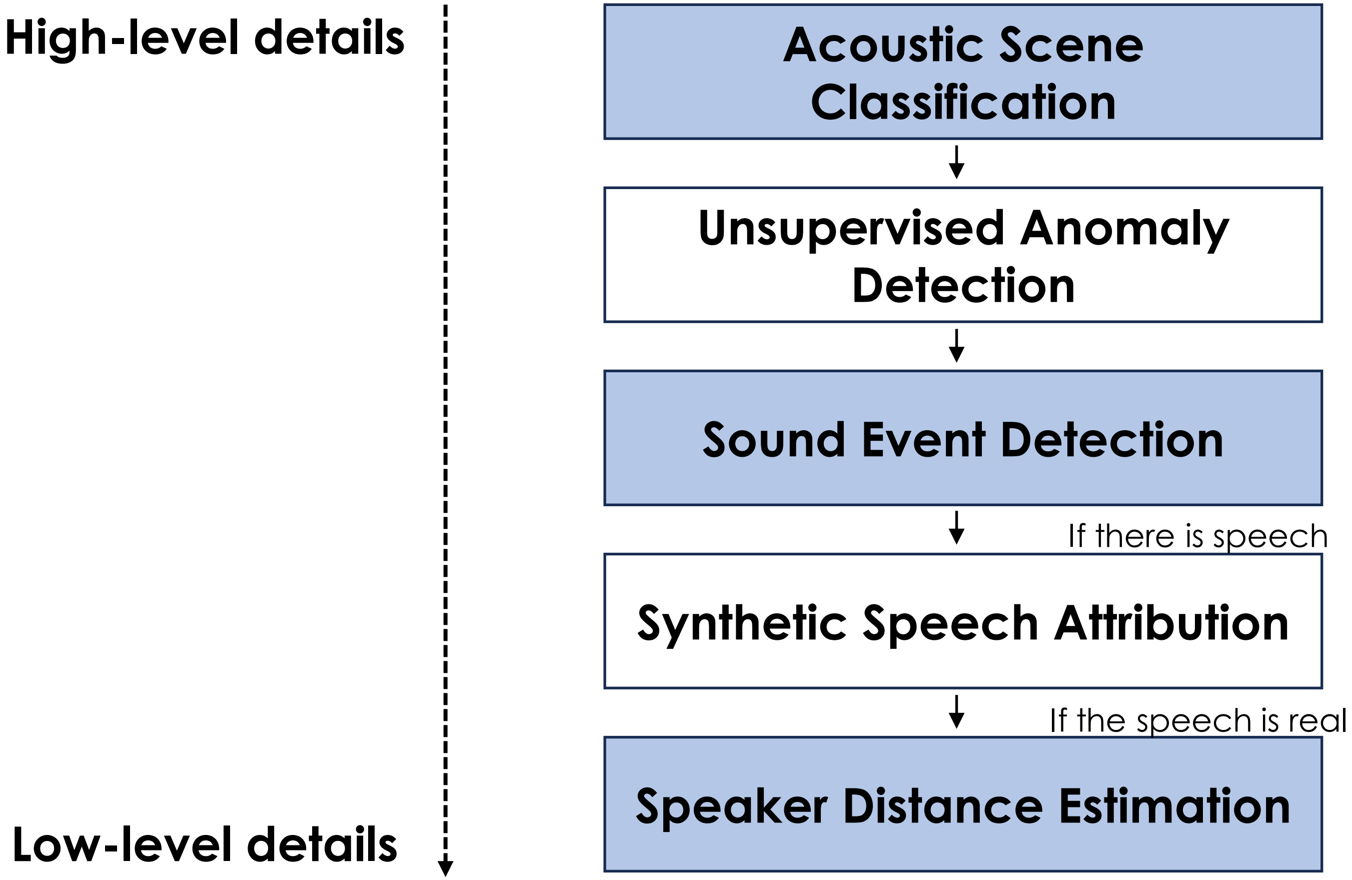
Main challenges

New explainable deep learning models

Low-complexity models for constrained environments

Reduction in training and inference time (real time)

Structure of the thesis



High-level details

Acoustic Scene Classification

Unsupervised Anomaly Detection

Sound Event Detection

If there is speech

Synthetic Speech Attribution

If the speech is real

Speaker Distance Estimation

Low-level details

"Where was the sound recorded?"

"Is the sound unusual with respect to the scene?"

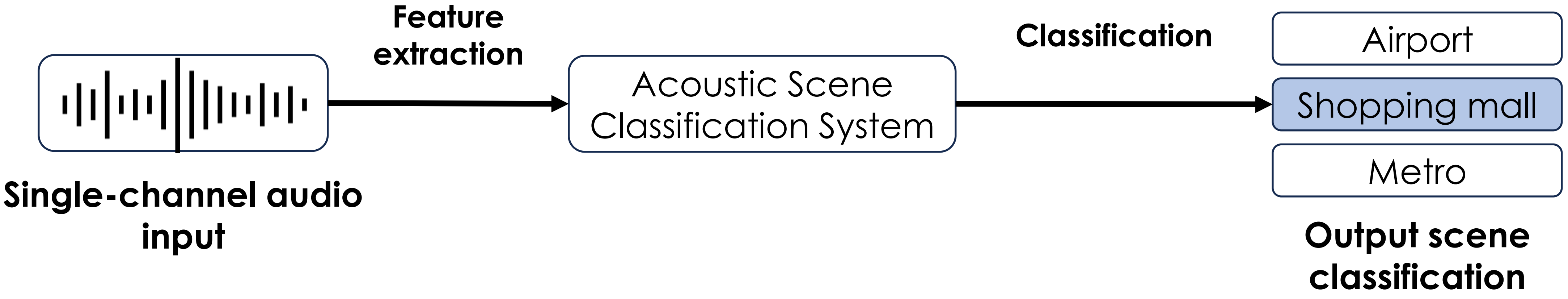
"What is happening in the scene?"
"When is happening?"

"Is the speech in the recording real?"

"How far is the speaker from the microphone?"

Acoustic Scene Classification

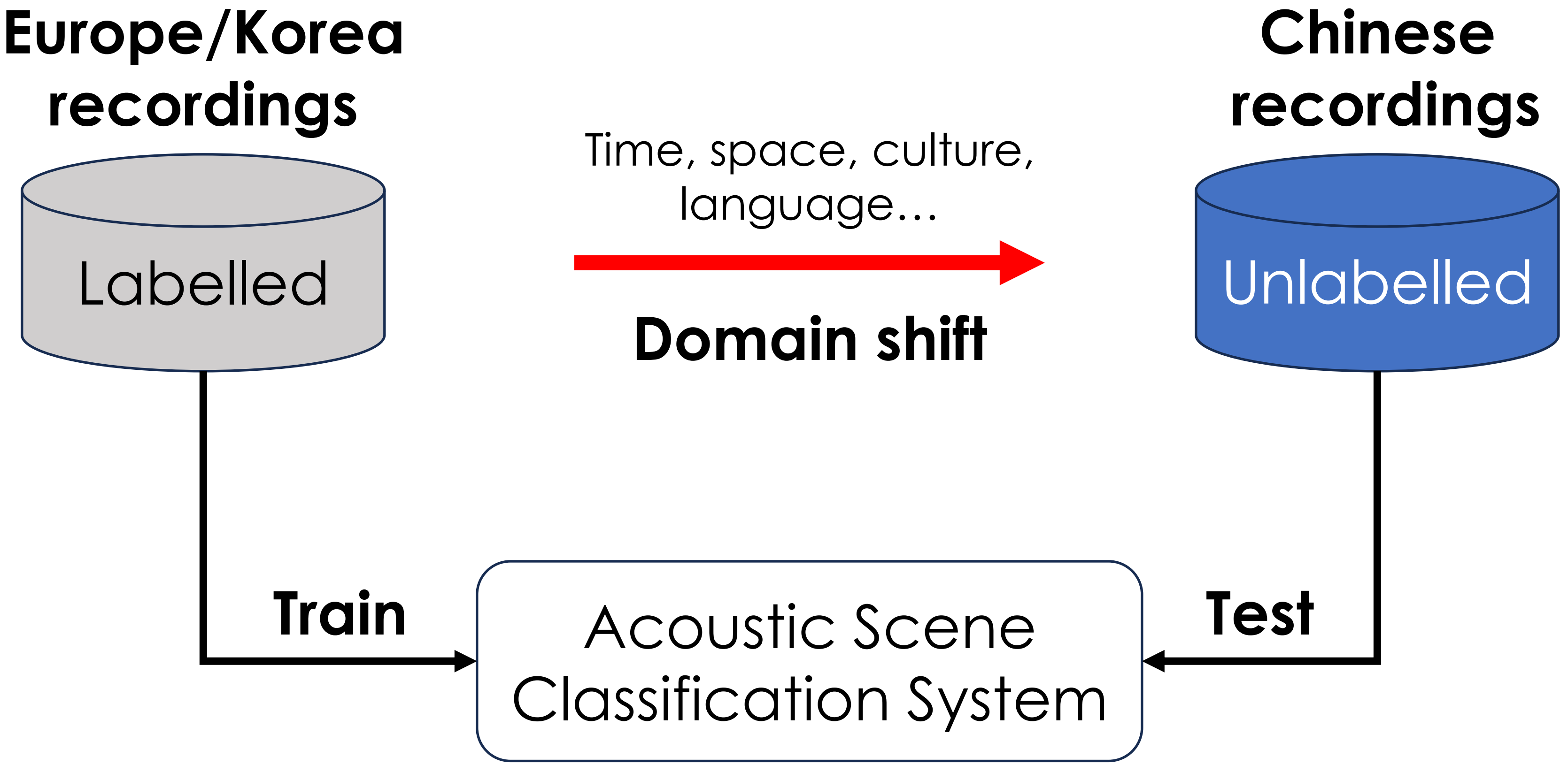
Acoustic scene classification (ASC) involves identifying an environment, such as an airport, shopping mall, or bus, based on **its acoustic characteristics and occurring events**.


Feature extraction
Classification
Airport
Shopping mall
Metro
Acoustic Scene Classification System
Single-channel audio input
Output scene classification

[C1] L. Pallotta, **M. Neri**, M. Buongiorno, A. Neri, and G. Giunta, "A Machine Learning-Based Approach for Audio Signals Classification using Chebyshev Moments and Mel-Coefficients", in: International Conference on Frontiers of Signal Processing (ICFSP), 2022.
[C2] M. Neri, L. Pallotta, and M. Carli, "Low-Complexity Environmental Sound Classification using Cadence Frequency Diagram and Chebyshev Moments", in: International Symposium on Image and Signal Processing and Analysis (ISPA), 2023.

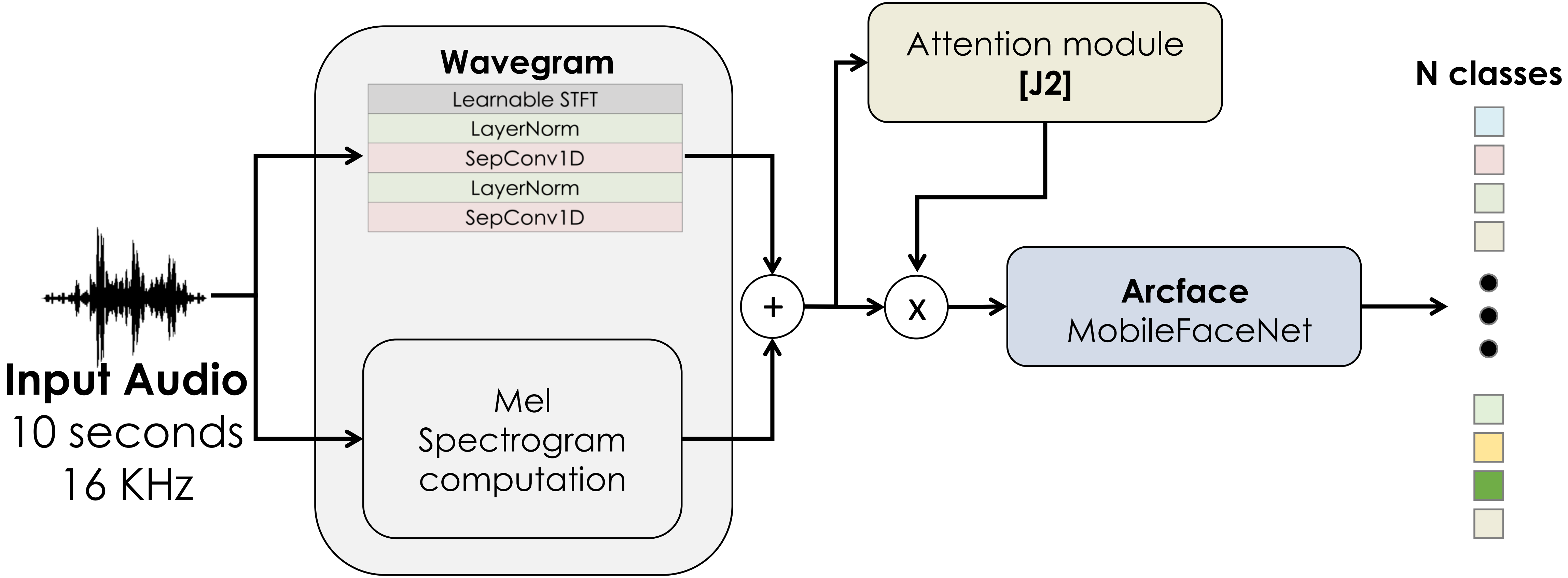
Domain Shift in ASC

Objective: Maximize accuracy on unseen recordings, improving robustness and generalization capabilities of a ASC model.



[C3] **M. Neri** and M. Carli, "Semi-Supervised Acoustic Scene Classification under Domain Shift using Attention-based Separable Convolutions and Angular Loss", in: IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2024.

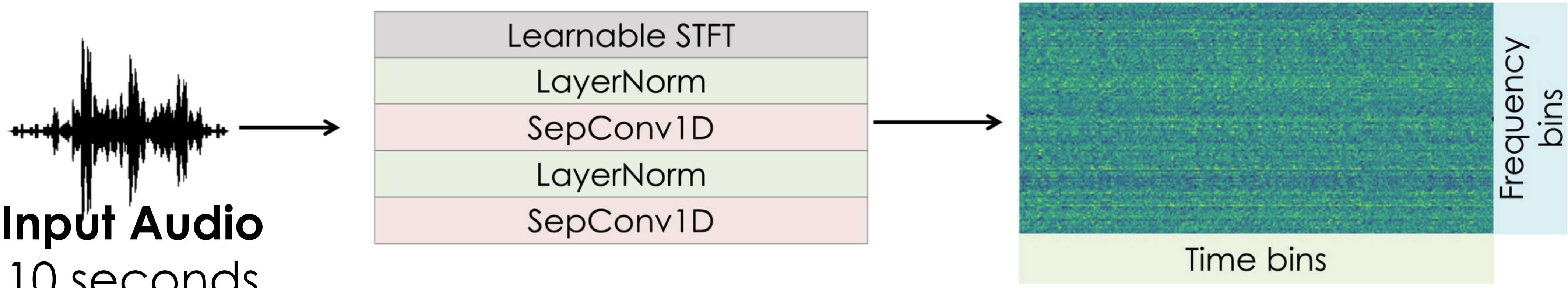
Proposed ASC



[C3] M. Neri and M. Carli, “Semi-Supervised Acoustic Scene Classification under Domain Shift using Attention-based Separable Convolutions and Angular Loss”, in: IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2024.
[J2] M. Neri, A. Politis, D. Krause, M. Carli, and T. Virtanen. “Speaker Distance Estimation in Enclosures from Single-Channel Audio”, in: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.
J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4685-4694, doi: 10.1109/CVPR.2019.00482.

Wavegram representation

To enhance the generalization capabilities of the approach, we introduce a low-complexity learnable filterbank called **Wavegram.**



Learnable STFT

LayerNorm

SepConv1D

LayerNorm

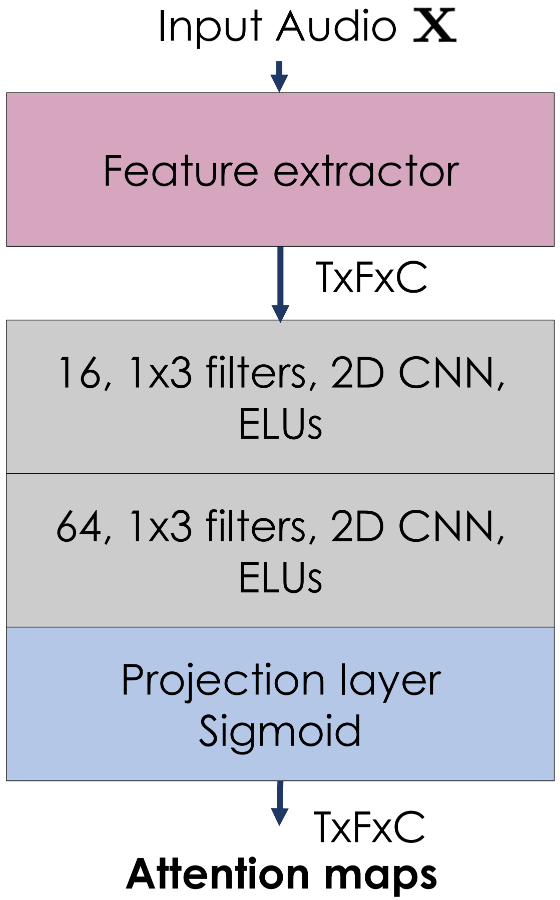
SepConv1D

Input Audio
10 seconds
16 KHz

Responsible for learning a **new time-frequency representation** in conjunction with the hand-crafted Mel-Spectrogram.

Our attention Module

A **convolutional neural network** is employed from [J2] for estimating the most salient time-frequency patterns from input audio $\mathbf{x}$.



Input Audio $\mathbf{X}$

Feature extractor

TxFxC

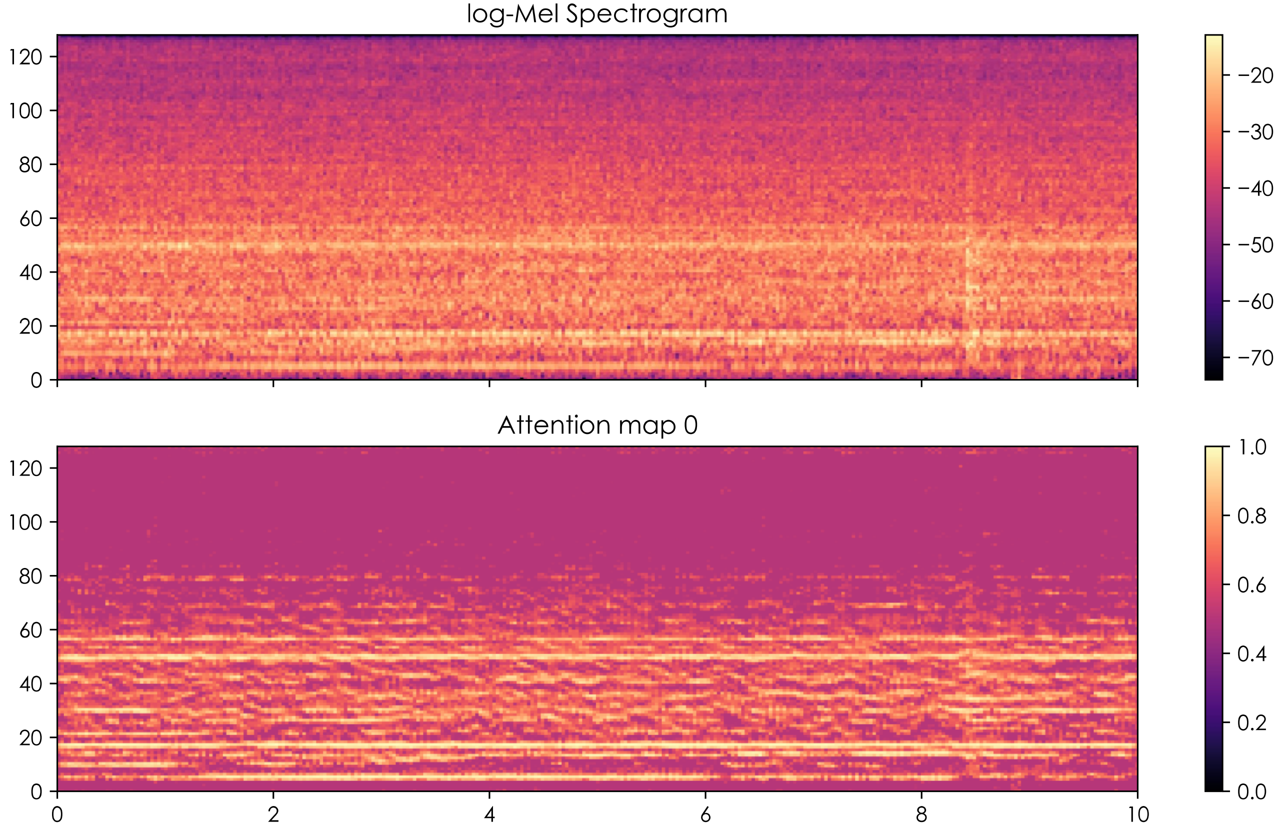
16, 1x3 filters, 2D CNN, ELUs

64, 1x3 filters, 2D CNN, ELUs

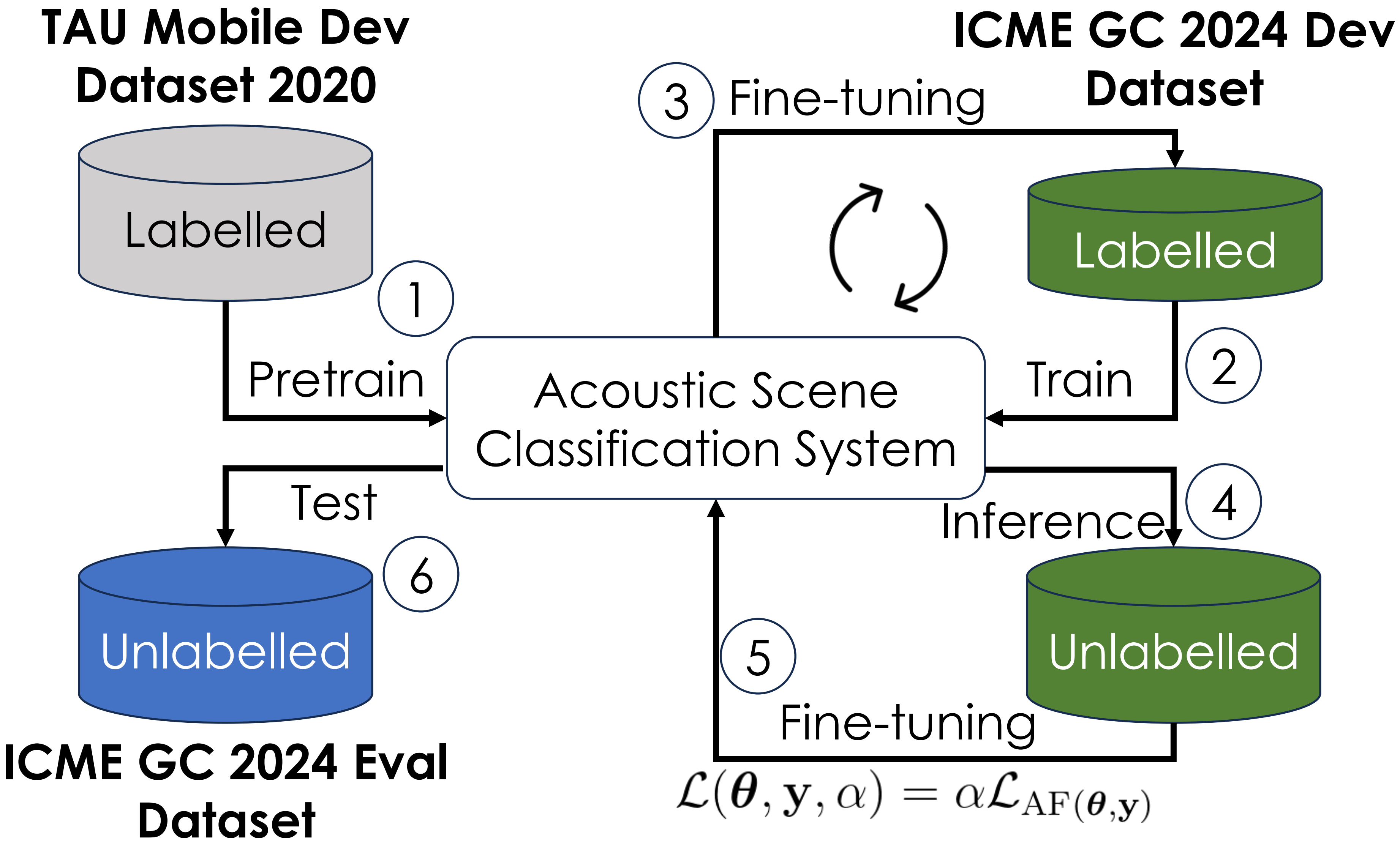
Projection layer Sigmoid

TxFxC

Attention maps



Semi-supervised algorithm



TAU Mobile Dev Dataset 2020

Labelled

1 Pretrain

ICME GC 2024 Dev Dataset

3 Fine-tuning

Labelled

2 Train

Acoustic Scene Classification System

4 Inference

Unlabelled

5 Fine-tuning

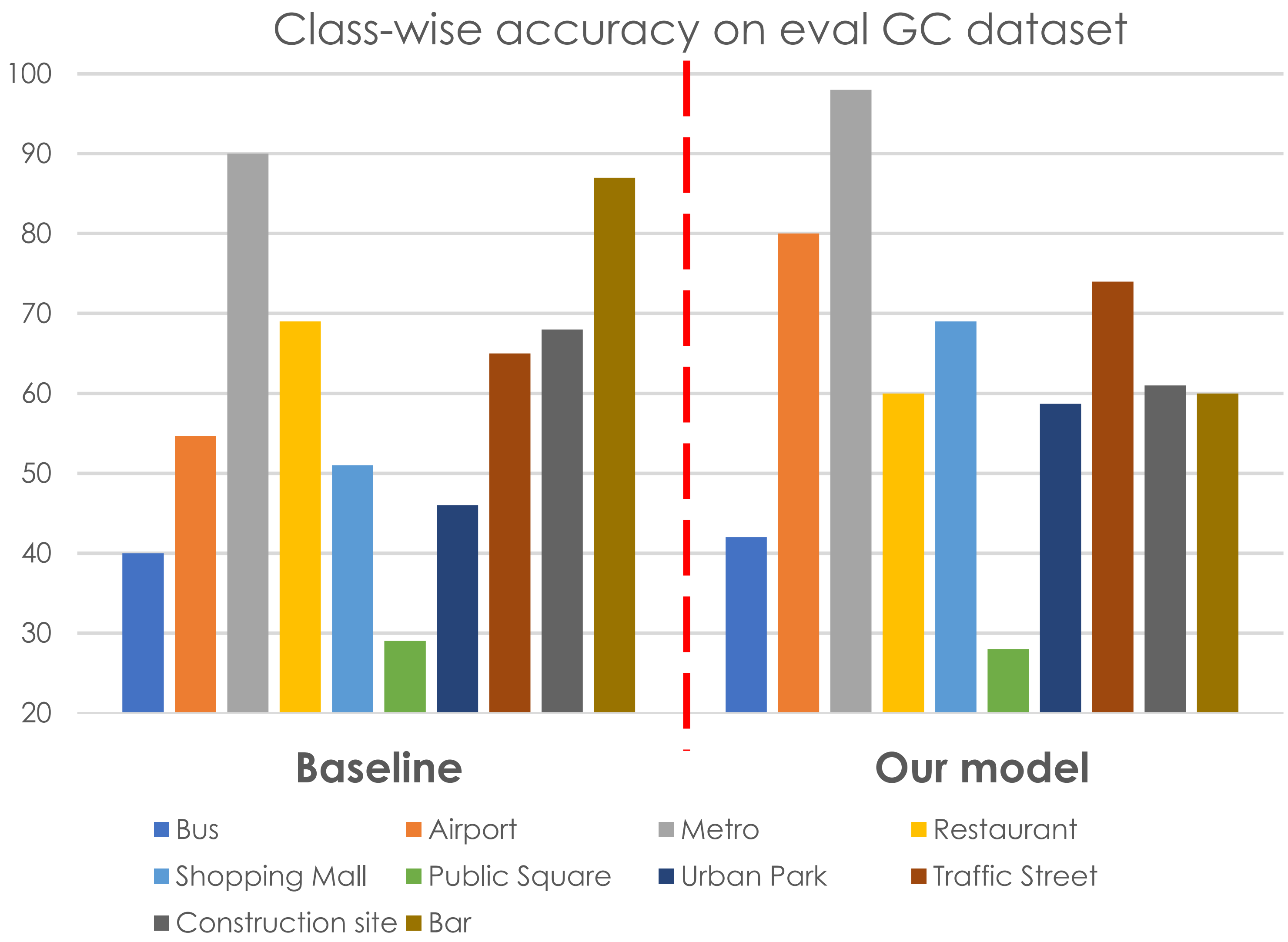
$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{y}, \alpha) = \alpha \mathcal{L}_{\mathrm{AF}(\boldsymbol{\theta}, \mathbf{y})}$$

Test 6

Unlabelled

ICME GC 2024 Eval Dataset

Test on evaluation dataset



Validation set performance

Approach	Macro-Accuracy (%)
No finetuning	63.8
1 fine-tuning iteration	97.7
2 fine-tuning iterations	98.3
3 fine-tuning iterations	**99.4**

Test set performance (blind submission)

- Baseline: 60%
- Ours: **63.1%**

Summary

- Definition of a new **semi-supervised algorithm** for accounting domain shift in acoustic scene classification.

- Successful combination of **ArcFace** and **Wavegram** for Acoustic Scene Classification.

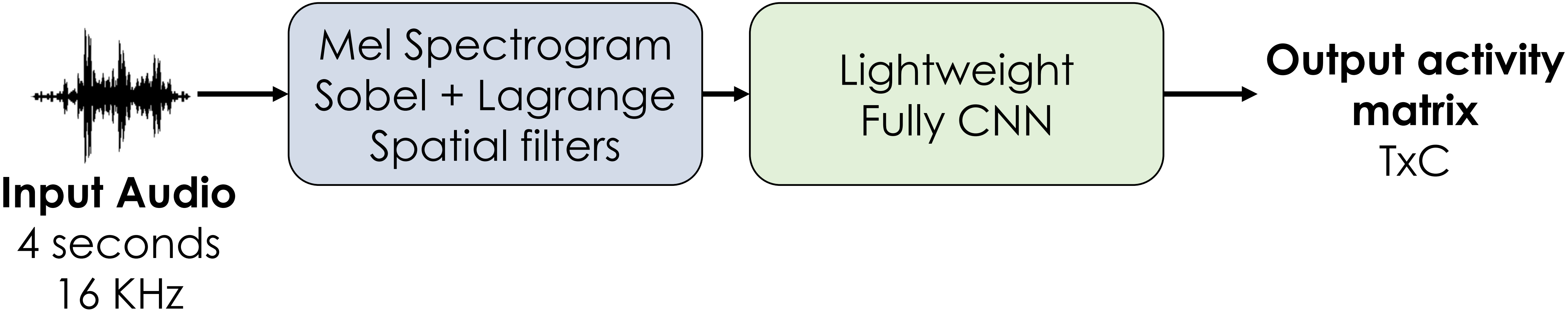
- **5th position** at the IEEE ICME 2024 **Grand Challenge** “Semi-supervised Acoustic Scene Classification under Domain Shift”, invited to **submit and present** at the conference.

- Still a **wide-open research problem** in the **machine learning-based multimedia processing** field when **data is scarce** from the **target domain**.

Sound Event Detection

Sound Event Detection (SED) is the task to recognize the occurring events from a recording, detailing onset and offset times.

Output: An **activity matrix of size TxC** (time frames and classes) that describes **what happened** in the scene.



Input Audio
4 seconds
16 KHz

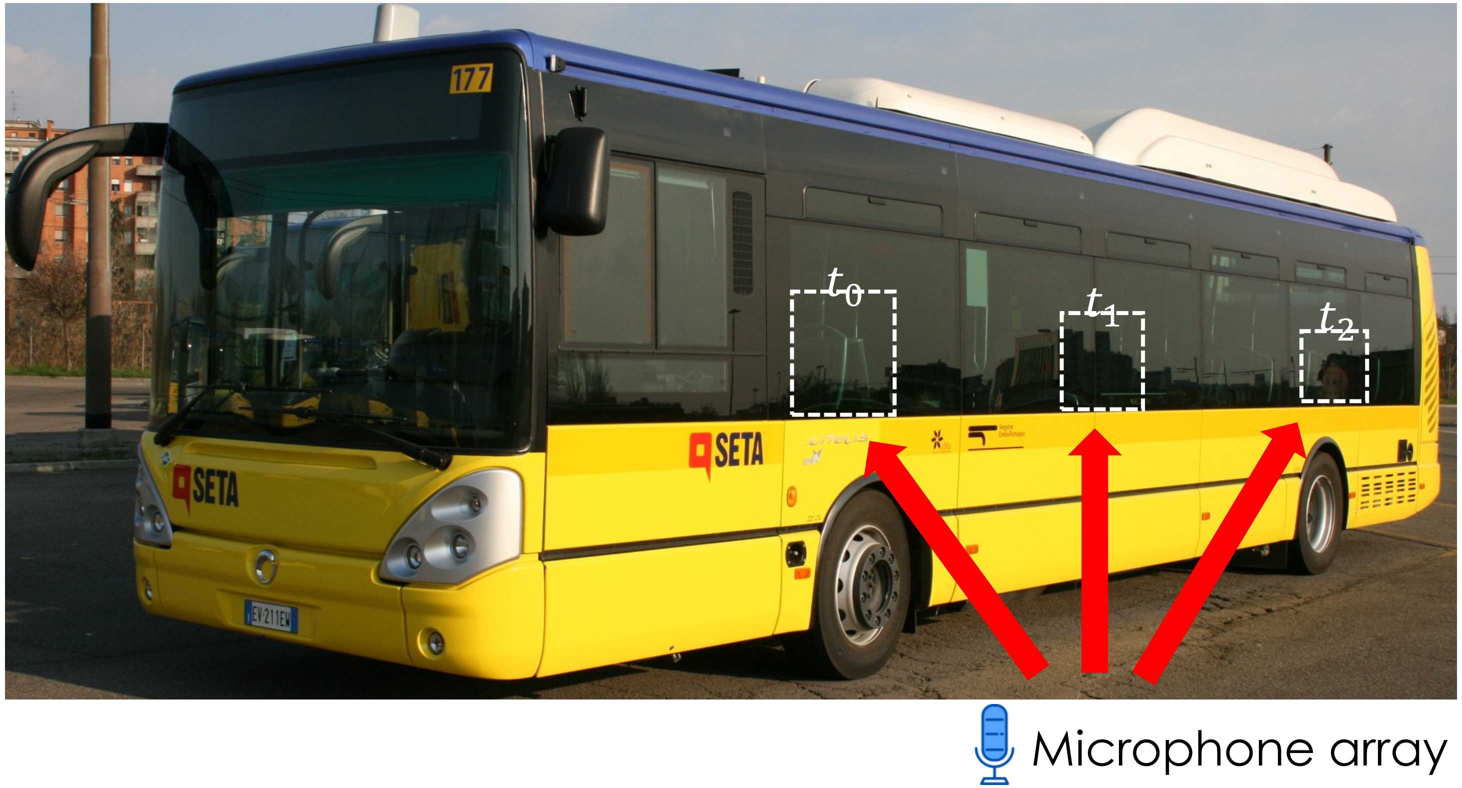
Mel Spectrogram
Sobel + Lagrange
Spatial filters

Lightweight
Fully CNN

Output activity matrix
TxC

[J3] M. Neri, F. Battisti, A. Neri, and M. Carli. "Sound Event Detection for Human Safety and Security in Noisy Environments", in: IEEE Access, 2022.

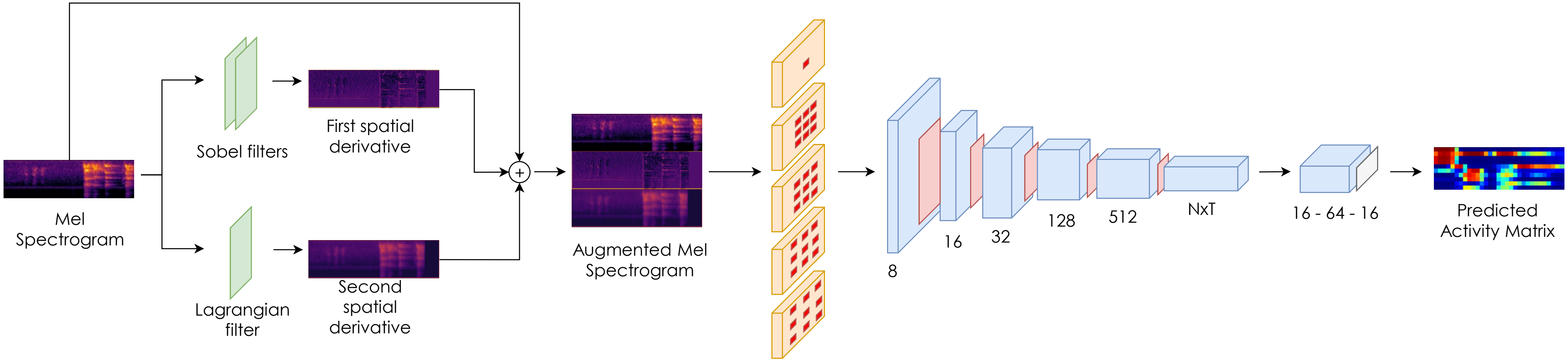
Collected SED Dataset



- **New dataset: SEDDOB (Sound Event Detection Dataset on Bus).**
- **Real noise** recordings from a bus.
- **10 anomalies** injected (e.g., gunshot, slap).
- Approx **11 hours** of labeled recordings with fine labels.

[J3] **M. Neri**, F. Battisti, A. Neri, and M. Carli. "Sound Event Detection for Human Safety and Security in Noisy Environments", in: IEEE Access, 2022.

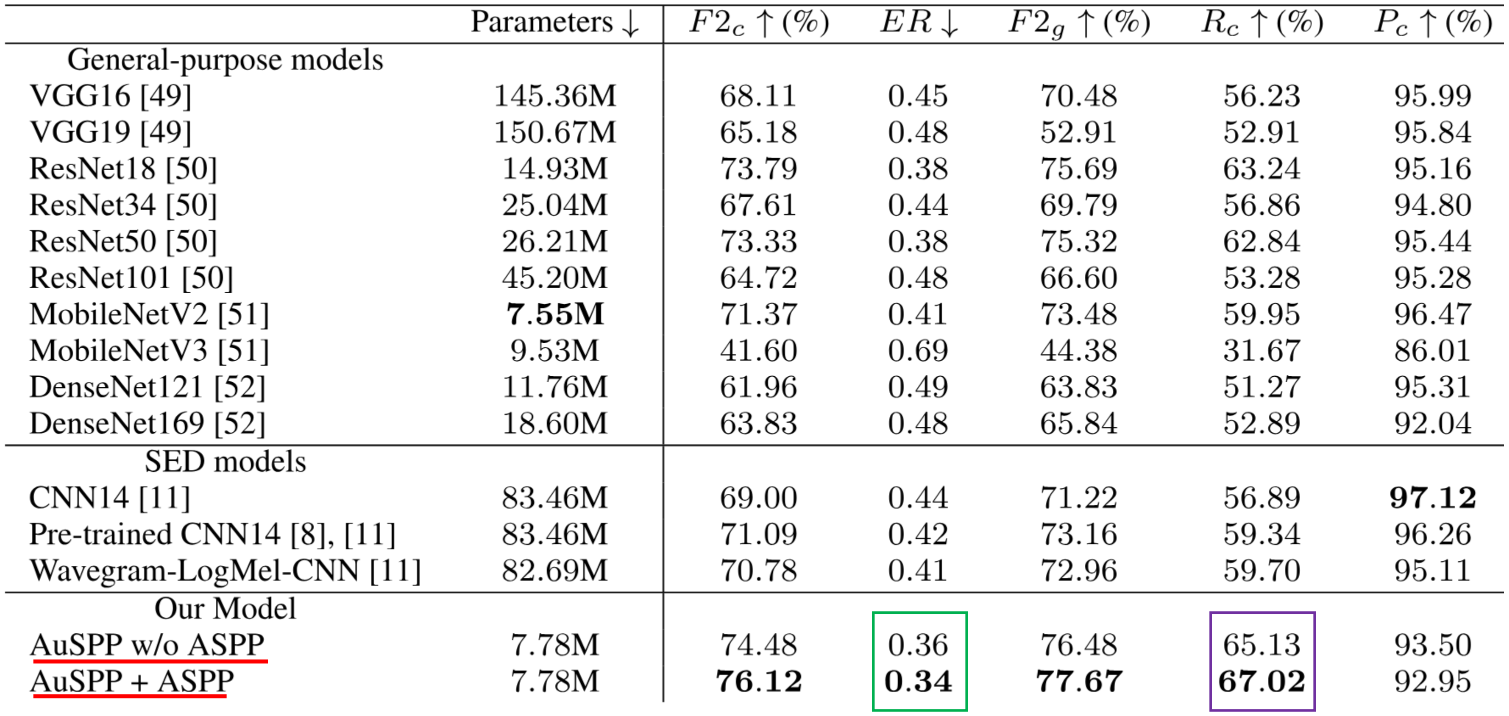
Proposed model: AuSPP



- **Pre-processing stage**: Spatial filters to compose the Augmented Mel Spectrogram.
- **Atrous Spatial Pyramid Pooling** applied to Augmented Mel Spectrogram.
- **Fully Convolutional Network** to predict the activity matrix.
- **Binary cross-entropy loss** between predicted and ground truth activity matrix.

[J3] M. Neri, F. Battisti, A. Neri, and M. Carli. “Sound Event Detection for Human Safety and Security in Noisy Environments”, in: IEEE Access, 2022.

Experimental results on SEDDOB



	Parameters ↓	$F2_c$ ↑ (%)	ER ↓	$F2_g$ ↑ (%)	R_c ↑ (%)	P_c ↑ (%)
General-purpose models						
VGG16 [49]	145.36M	68.11	0.45	70.48	56.23	95.99
VGG19 [49]	150.67M	65.18	0.48	52.91	52.91	95.84
ResNet18 [50]	14.93M	73.79	0.38	75.69	63.24	95.16
ResNet34 [50]	25.04M	67.61	0.44	69.79	56.86	94.80
ResNet50 [50]	26.21M	73.33	0.38	75.32	62.84	95.44
ResNet101 [50]	45.20M	64.72	0.48	66.60	53.28	95.28
MobileNetV2 [51]	**7.55M**	71.37	0.41	73.48	59.95	96.47
MobileNetV3 [51]	9.53M	41.60	0.69	44.38	31.67	86.01
DenseNet121 [52]	11.76M	61.96	0.49	63.83	51.27	95.31
DenseNet169 [52]	18.60M	63.83	0.48	65.84	52.89	92.04
SED models						
CNN14 [11]	83.46M	69.00	0.44	71.22	56.89	**97.12**
Pre-trained CNN14 [8], [11]	83.46M	71.09	0.42	73.16	59.34	96.26
Wavegram-LogMel-CNN [11]	82.69M	70.78	0.41	72.96	59.70	95.11
Our Model						
AuSPP w/o ASPP	7.78M	74.48	0.36	76.48	65.13	93.50
AuSPP + ASPP	7.78M	**76.12**	**0.34**	**77.67**	**67.02**	92.95

- Performance obtained with **confidence threshold** > 80%, 50ms of time segments.

- Best **Error Rate** and **Recall** with respect to SOTA approaches

- **Lowest number of learnable parameters** among SOTA sound event detectors.

[J3] **M. Neri**, F. Battisti, A. Neri, and M. Carli. “Sound Event Detection for Human Safety and Security in Noisy Environments”, in: IEEE Access, 2022.

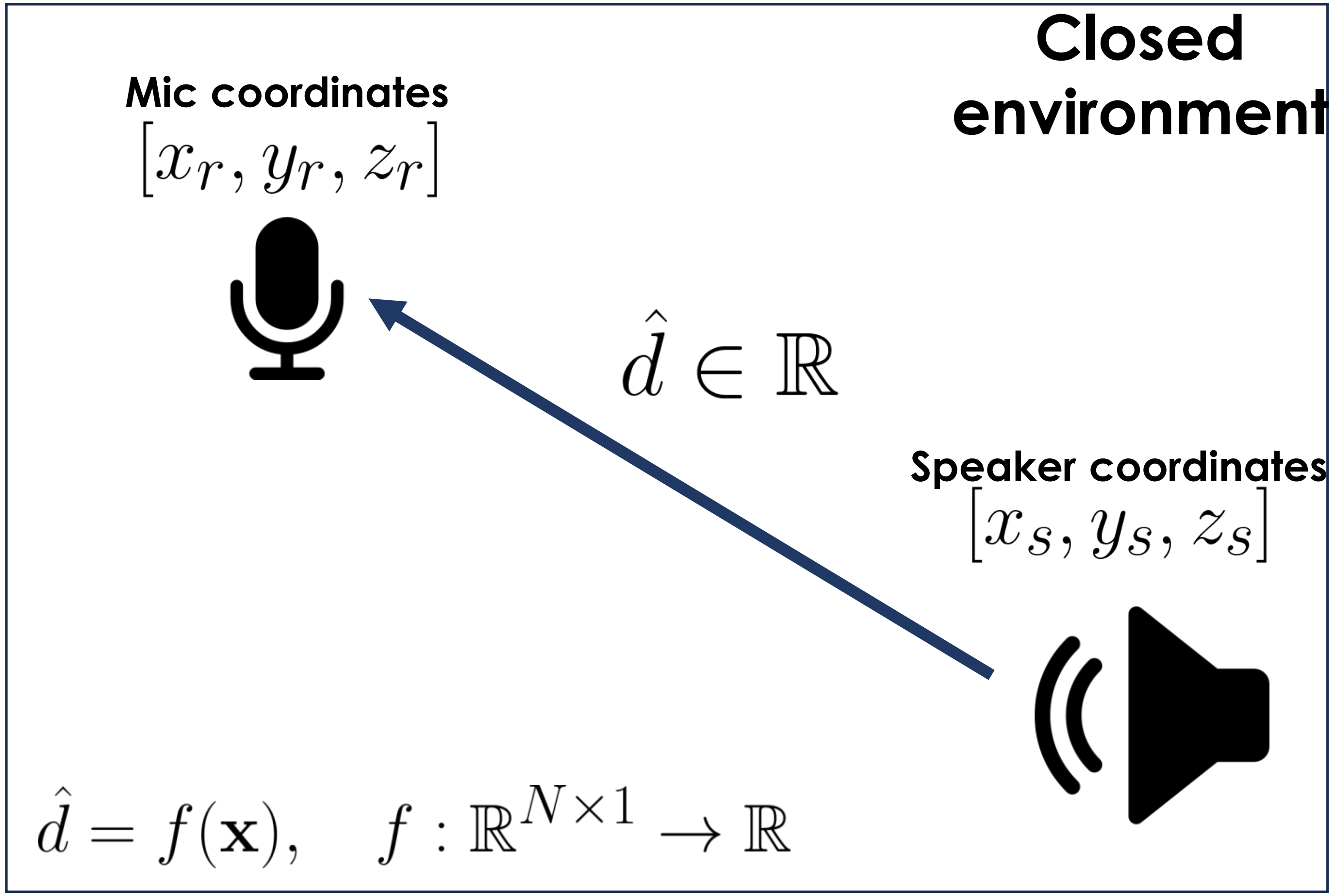
Summary

- Release of a new **Sound Event Detection dataset (SEDDOB)** in which **real tram background noise has been collected** for the first time in literature. Anomaly events has been synthetically injected.

- Definition of a new **low-complexity Sound Event Detector (AuSPP)** that simultaneously **capture short and long-term time-frequency dependencies from the Augmented Mel-spectrograms**.

- Comparison with SOTA approaches demonstrates the **superiority of AuSPP, both in terms of recognition performance and complexity**.

Speaker Distance Estimation



Closed environment

Mic coordinates

$[x_r, y_r, z_r]$

$\hat{d} \in \mathbb{R}$

Speaker coordinates

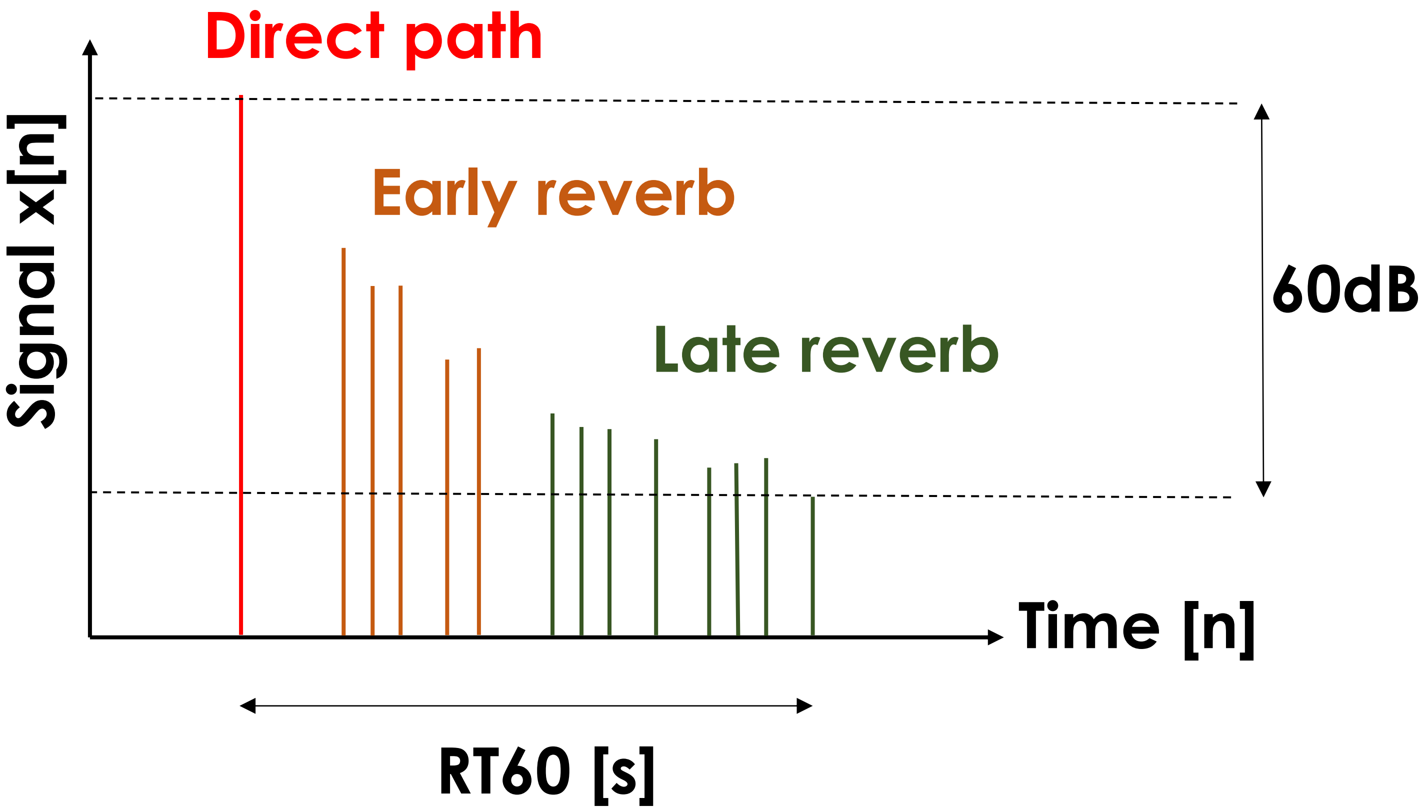
$[x_s, y_s, z_s]$

$$\hat{d} = f(\mathbf{x}), \quad f : \mathbb{R}^{N \times 1} \rightarrow \mathbb{R}$$

- **New task**: estimate a real-valued distance between microphone and speaker, without knowing their locations.
- **No room acoustic characteristics.**
- **No guidelines** regarding **features for distance estimation** from single-channel recordings.

[C5] **M. Neri**, A. Politis, D. Krause, M. Carli, and T. Virtanen, "Single-Channel Speaker Distance Estimation in Reverberant Environments", in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2023.

Recall: Reverberation cues

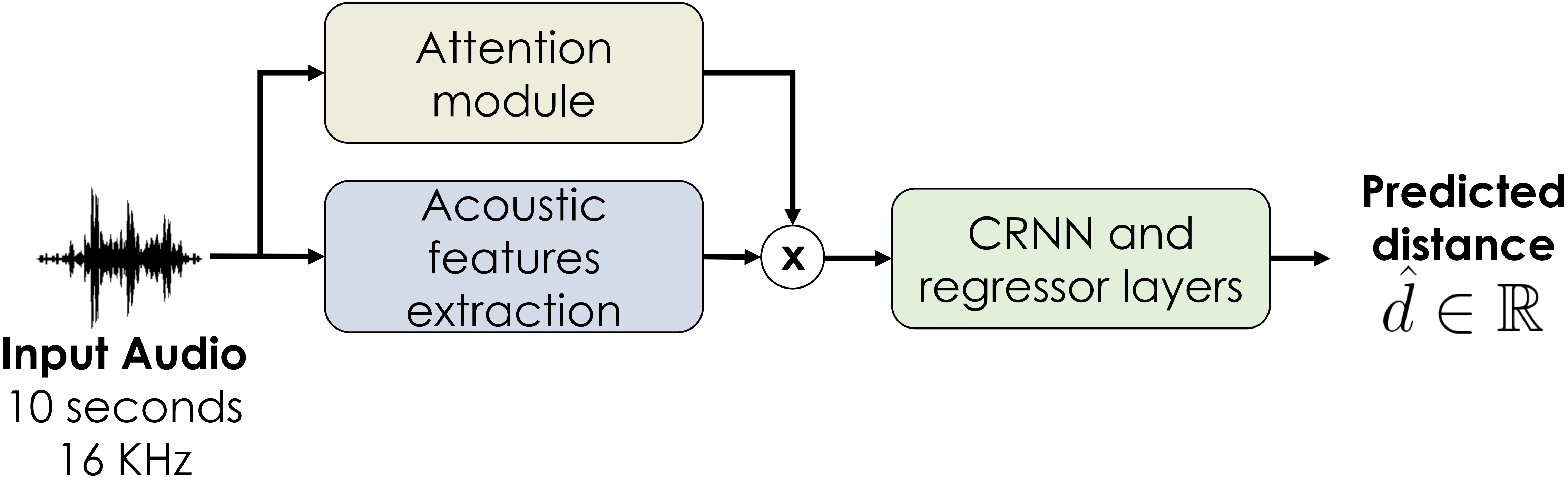


- Without multiple mics, **reverberation can be exploited** from single-channel recordings.

- **Reverb is strictly correlated with room acoustics** (multipath phenomena).

- Implicitly modelled by a **Deep Neural Network**.

Speaker Distance Estimation



- Fully convolutional attention network.
- Acoustic features: **STFT magnitude** and **sin&cos of STFT phase**.
- Two Gated Recurrent Units with MultiLayer Perceptron regressors.

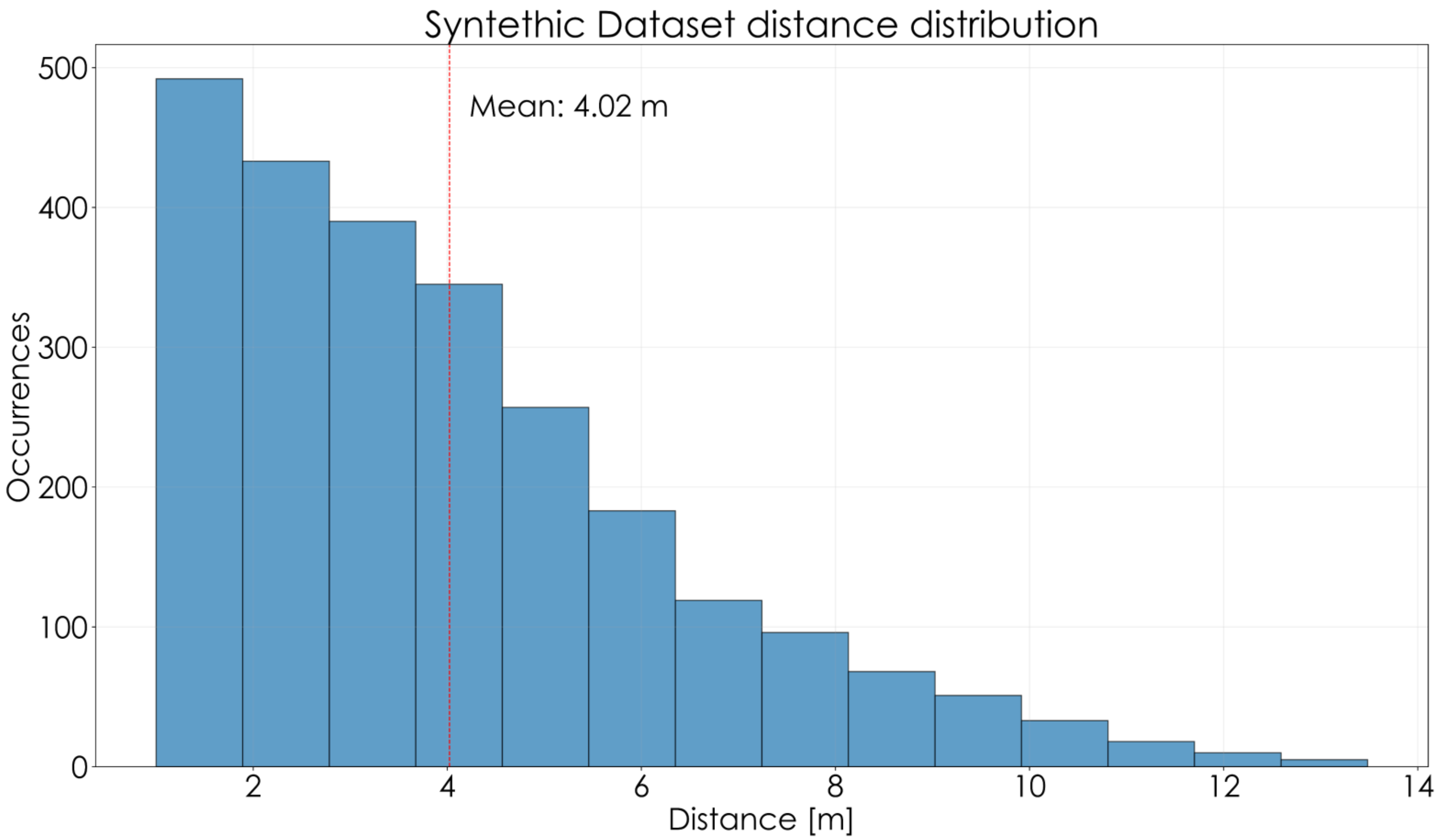
[J2] M. Neri, A. Politis, D. Krause, M. Carli, and T. Virtanen. "Speaker Distance Estimation in Enclosures from Single-Channel Audio", in: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

Synthetic Dataset Generation

Parameter	Random ranges
Room width and length	[3.0, 15.0] m
Room height	[2.0, 7.0] m
Number of materials (wall, floor, ceiling)	13, 7, 8
Source – receiver height	[1.5, 2.2] m
Source-to-surface distance	> 0.5 m
Source-to-receiver distance	> 1.0 m

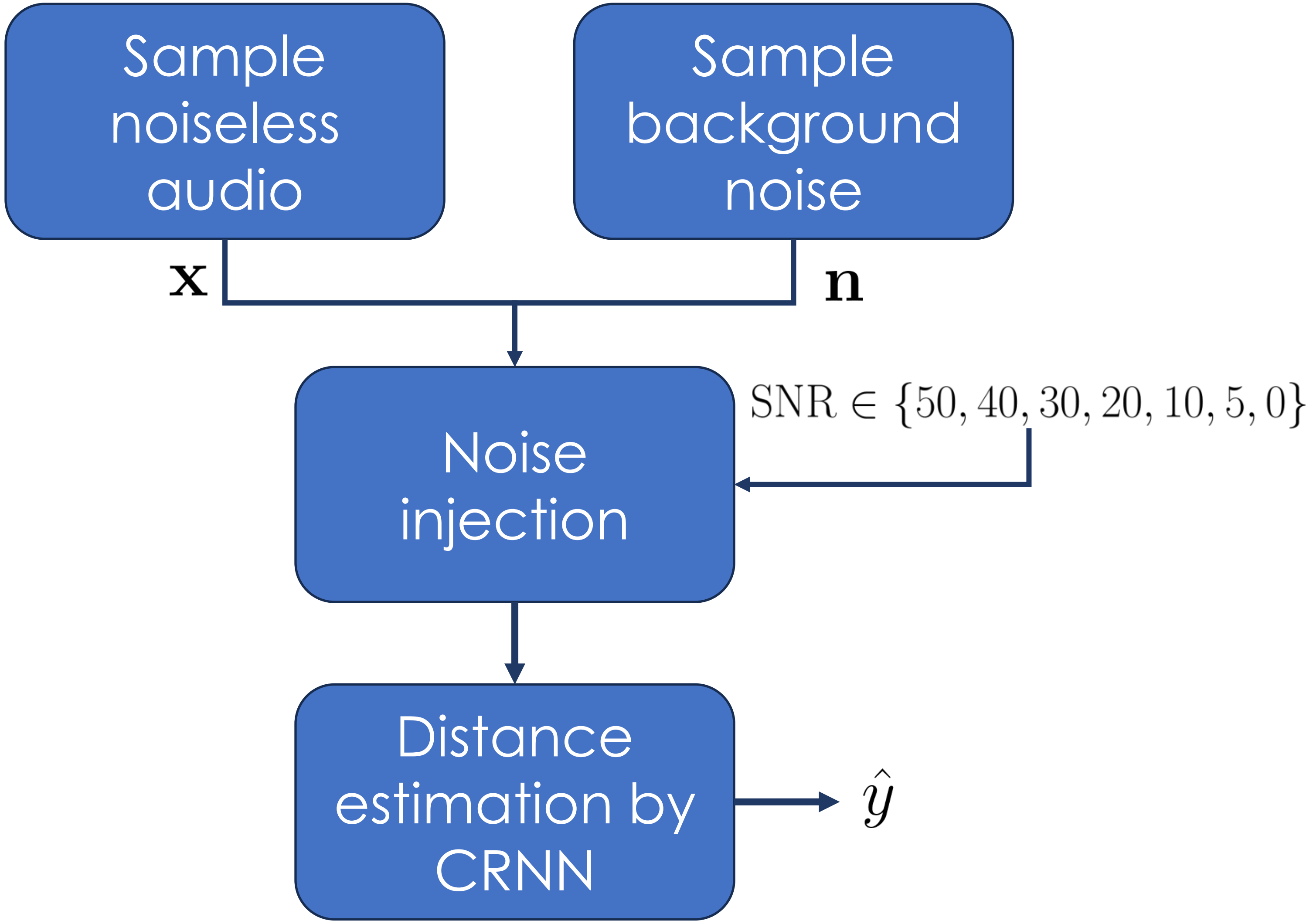
- **Image-source simulator** for **Room Impulse Responses** generation and recording synthesis.
- Anechoic speech from **TIMIT dataset**.
- **2500 different rooms** split into 5 folds for computing confidence intervals.

Synthetic Dataset Generation



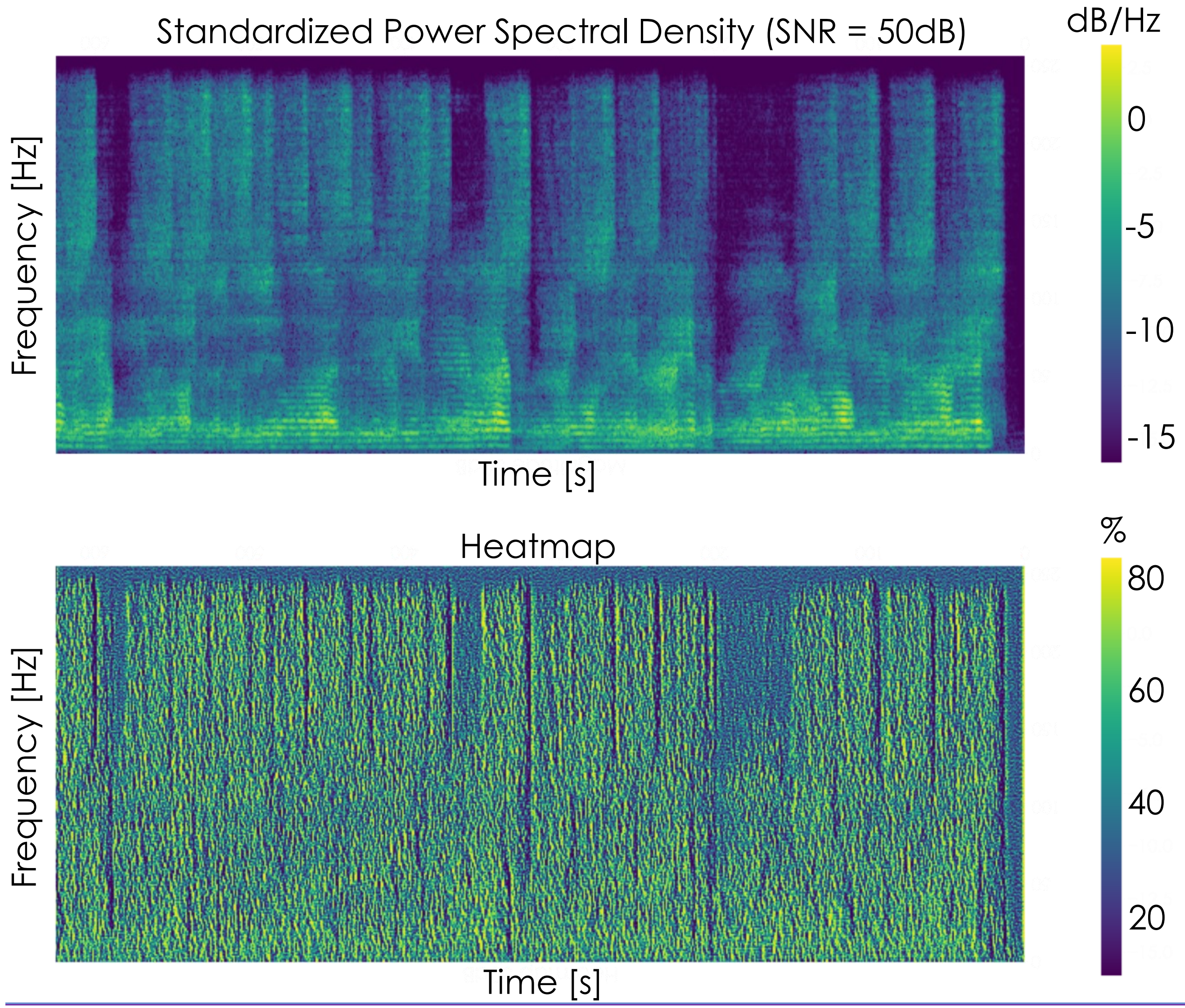
- **Unique simulated rooms** with their r**oom-source-distance** configuration.
- **Unbalanced dataset** (low number of samples for distant speaker) **but representative of real scenarios.**

Synthetic Dataset Generation



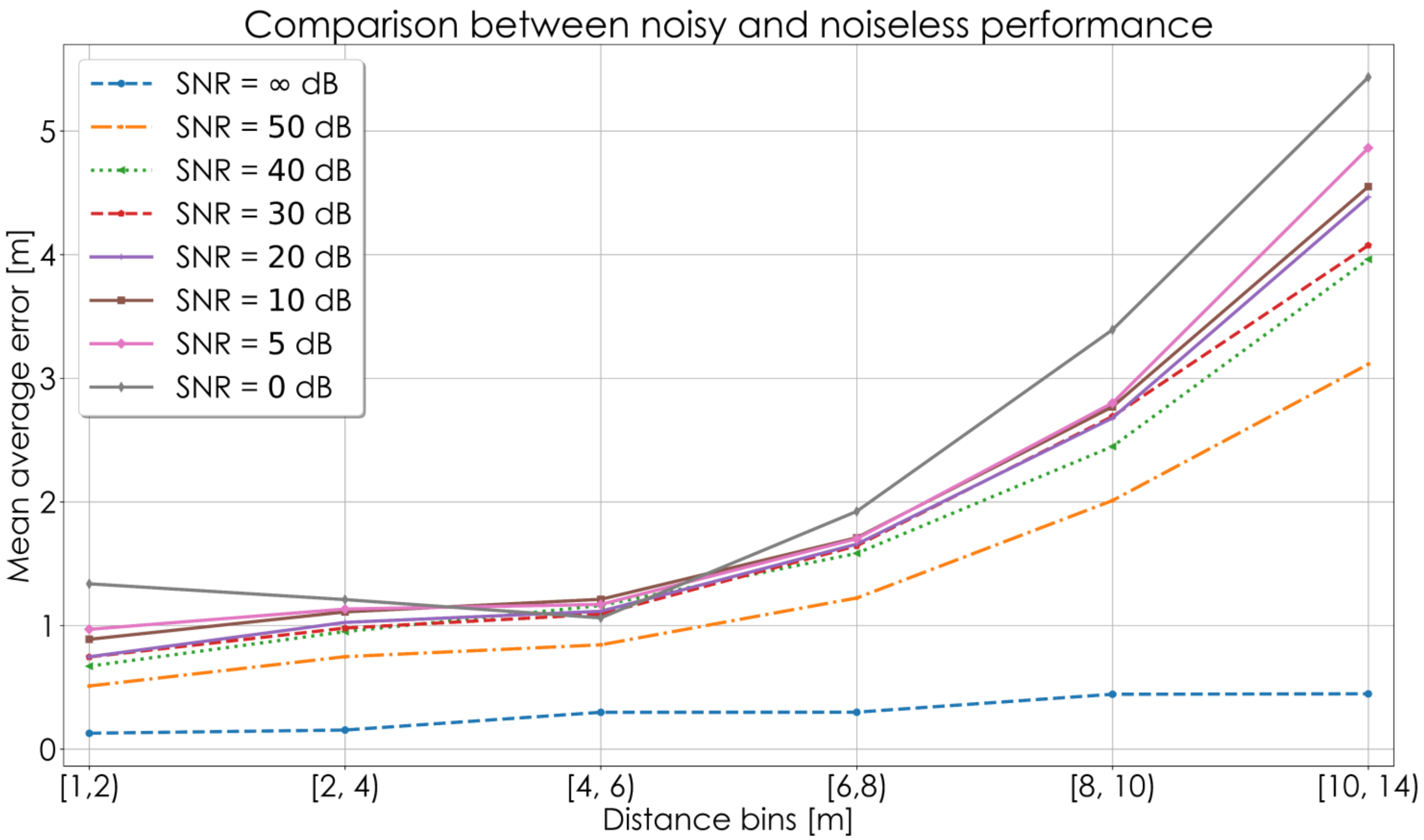
- Noisy version of the dataset by injecting **real environmental recordings**
- WHAMR! Dataset with high quality noisy backgrounds (restaurants, café, public transportations)
- Different noises in training, validation, and testing

Results on noiseless synthetic data



- Attention module for explainability.
- Distance estimation error of around **10 cm** in the synthetic scenario without noise.
- **Not only speech harmonics** (up to 4kHz) **are useful** for the task.

Results on noisy synthetic data



- **Large discrepancy** between **noiseless and noisy results**.
- **Phase-based features** are severely **corrupted** by tiny amount of noise.
- **Loss of phase coherence** across frequencies due to noise.

Results on hybrid and real datasets

Cross-dataset w/o fine-tuning

	Synthetic	Hybrid	Real
Synthetic	0.11 ± 0.00	4.28 ± 0.45	4.14 ± 0.08
Hybrid	6.80 ± 0.59	1.52 ± 0.12	3.76 ± 0.56
Real	2.26 ± 0.38	8.22 ± 0.54	0.42 ± 0.02

Cross-dataset w/ fine-tuning

	Synthetic	Hybrid	Real
Synthetic	0.11 ± 0.00	1.57 ± 0.23	0.47 ± 0.05
Hybrid	0.18 ± 0.04	1.52 ± 0.12	0.45 ± 0.05
Real	0.11 ± 0.02	1.54 ± 0.22	0.42 ± 0.02

- **Hybrid:** recorded Room Impulse Response convolved with synthetic speech (QMULTIMIT).

- **Real:** on-the-field recordings (STARSS23).

- **High error across different data sources.** With enough target data, domain shift can be mitigated.

Summary

- Introduction of a new challenge: *continuous-valued* **speaker distance estimation**.

- It is possible to estimate the distance between speaker and the microphone **without the use of multi-channel recordings** with **low-complexity neural networks**.

- **The nature of Room Impulse Response** (synthetic-real) **can cause domain shift** in this task. ICASSP 2025 GENDA workshop organized a challenge on this problem.

- The definition of the attention module enabled to investigate **model predictions and interpret most salient time-frequency patterns** for the resolution of the task.

Final Remarks

Other Academic Contributions

Managing Editor for *Signal Processing: Image Communications* Elsevier journal from September 2024.

Reviewer for several journals and conferences:

- **Journals**. *IEEE Transactions on Image Processing*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Broadcasting*, *Expert Systems with Application*, *Signal Processing: Image Communications*, *IEEE Access*.
- **Conferences**. IEEE International Conference on Multimedia & Expo (ICME), IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), IEEE International Workshop on Multimedia Signal Processing (MMSP), Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), IEEE International Workshop on Information Forensics and Security (WIFS), International Symposium on Image and Signal Processing and Analysis (ISPA).

In 2023 I was Local Arrangement Co-Chair for ISPA and a Session Chair (*Visual Data Acquisition and Computation Session*) for IEEE ICME 2024.

I co-supervised 5 M.Sc. thesis and 2 B.Sc. thesis.

List of Publications

Journal articles

[J1] M. Neri and M. Carli. "Low-complexity Unsupervised Audio Anomaly Detection exploiting Separable Convolutions and Angular Loss", in: **IEEE Sensors Letters**, 2024.

[J2] M. Neri, A. Politis, D. Krause, M. Carli, and T. Virtanen. "Speaker Distance Estimation in Enclosures from Single-Channel Audio", in: **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, 2024.

[J3] M. Neri, F. Battisti, A. Neri, and M. Carli. "Sound Event Detection for Human Safety and Security in Noisy Environments", in: **IEEE Access**, 2022.

[J4] K. Lamichhane, **M. Neri**, P. Pradip, F. Battisti, and M. Carli, "No-Reference Light Field Image Quality Assessment Exploiting Saliency", in: **IEEE Transactions on Broadcasting**, 2023.

[J5] M. Neri, "Anomaly Detection and Classification of Audio Signals with Artificial Intelligence Techniques", in: **Science Talks**, 2024.

[J6] M. Bernabei, S. Colabianchi, M. Carli, F. Costantino, A. Ferrarotti, **M. Neri**, S. Stabile, "Enhancing occupational safety and health training: a guideline for virtual reality integration", in: **IEEE Access**, 2024.

[J7] M. Neri and F. Battisti, "Low-Complexity Patch-based No-Reference Point Cloud Quality Metric exploiting Weighted Structure and Texture Features", in: **IEEE Transactions on Broadcasting**, 2025.

[J8] M. Neri and T. Virtanen, "Multi-channel Replay Speech detection using an Adaptive Learnable Beamformer", in: **IEEE Open Journal of Signal Processing**, 2025.

List of Publications

Conferences articles

[C1] L. Pallotta, **M. Neri**, M. Buongiorno, A. Neri, and G. Giunta, “A Machine Learning-Based Approach for Audio Signals Classification using Chebychev Moments and Mel-Coefficients”, in: International Conference on Frontiers of Signal Processing (ICFSP), 2022.

[C2] M. Neri, L. Pallotta, and M. Carli, “Low-Complexity Environmental Sound Classification using Cadence Frequency Diagram and Chebychev Moments”, in: International Symposium on Image and Signal Processing and Analysis (ISPA), 2023.

[C3] M. Neri and M. Carli, “Semi-Supervised Acoustic Scene Classification under Domain Shift using Attention-based Separable Convolutions and Angular Loss”, in: IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2024.

[C4] M. Neri, A. Ferrarotti, L. de Luisa, A. Salimbeni, and M. Carli, “ParalMGC: Multiple Audio Representations for Synthetic Human Speech Attribution”, in: 10th European Workshop on Visual Information Processing (EUVIP), 2022.

[C5] M. Neri, A. Politis, D. A. Krause, M. Carli, and T. Virtanen, “Single-Channel Speaker Distance Estimation in Reverberant Environments”, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2023.

[C6] S. Baldoni, F. Battisti, M. Brizzi, **M. Neri**, and A. Neri, “A Semantic Segmentation-based Approach for Train Positioning”, in: ITM/PTTI Institute Of Navigation (ION), 2022.

[C7] M. Neri and F. Battisti, “3D Object Detection on Synthetic Point Clouds for Railway Applications”, in: 10th European Workshop on Visual Information Processing (EUVIP), 2022.

[C8] M. Neri and M. Carli, “Artificial Intelligence Techniques for Quality Assessments of Immersive Multimedia”, in: ACM International Conference on Interactive Media Experiences (IMX), 2023.

List of Publications

Conferences articles

[C9] R. Bentivenga, M. Bernabei, M. Carli, S. Colabianchi, F. Costantino, A. Ferrarotti, **M. Neri**, E. Pietrafesa, E. Sorrentino, S. Stabile, “Advancing Occupational Safety and Health training: a Safety-II integration of the ADDIE model for virtual reality”, in: Methodologies and Intelligent Systems for Technology Enhanced Learning, 14th International Conference, 2024.

[C10] R. Bentivenga, M. Bernabei, M. Carli, S. Colabianchi, F. Costantino, A. Ferrarotti, **M. Neri**, E. Pietrafesa, E. Sorrentino, S. Stabile, “Transforming Training With New Enabling Technologies: A Proposal To Verify The Efficacy Of Virtual Reality Tools In The Occupational Health And Safety Sector”, in: 8th World Conference on Smart Trends in systems, Security, and Sustainability (Worlds4), 2024.

Thank you for the attention!
Questions?

Contacts
michael.neri@tuni.fi
github.com/michaelneri
michaelneri.github.io

List of Publications

Journal articles

[J1] M. Neri and M. Carli. “Low-complexity Unsupervised Audio Anomaly Detection exploiting Separable Convolutions and Angular Loss”, in: **IEEE Sensors Letters**, 2024.

[J2] M. Neri, A. Politis, D. Krause, M. Carli, and T. Virtanen. “Speaker Distance Estimation in Enclosures from Single-Channel Audio”, in: **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, 2024.

[J3] M. Neri, F. Battisti, A. Neri, and M. Carli. “Sound Event Detection for Human Safety and Security in Noisy Environments”, in: **IEEE Access**, 2022.

[J4] K. Lamichhane, **M. Neri**, P. Pradip, F. Battisti, and M. Carli, “No-Reference Light Field Image Quality Assessment Exploiting Saliency”, in: **IEEE Transactions on Broadcasting**, 2023.

[J5] M. Neri, “Anomaly Detection and Classification of Audio Signals with Artificial Intelligence Techniques”, in: **Science Talks**, 2024.

[J6] M. Bernabei, S. Colabianchi, M. Carli, F. Costantino, A. Ferrarotti, **M. Neri**, S. Stabile, “Enhancing occupational safety and health training: a guideline for virtual reality integration”, in: **IEEE Access**, 2024.

[J7] M. Neri and F. Battisti, “Low-Complexity Patch-based No-Reference Point Cloud Quality Metric exploiting Weighted Structure and Texture Features”, in: **IEEE Transactions on Broadcasting**, 2025.

[J8] M. Neri and T. Virtanen, “Multi-channel Replay Speech detection using an Adaptive Learnable Beamformer”, in: **IEEE Open Journal of Signal Processing**, 2025.