UC San Diego

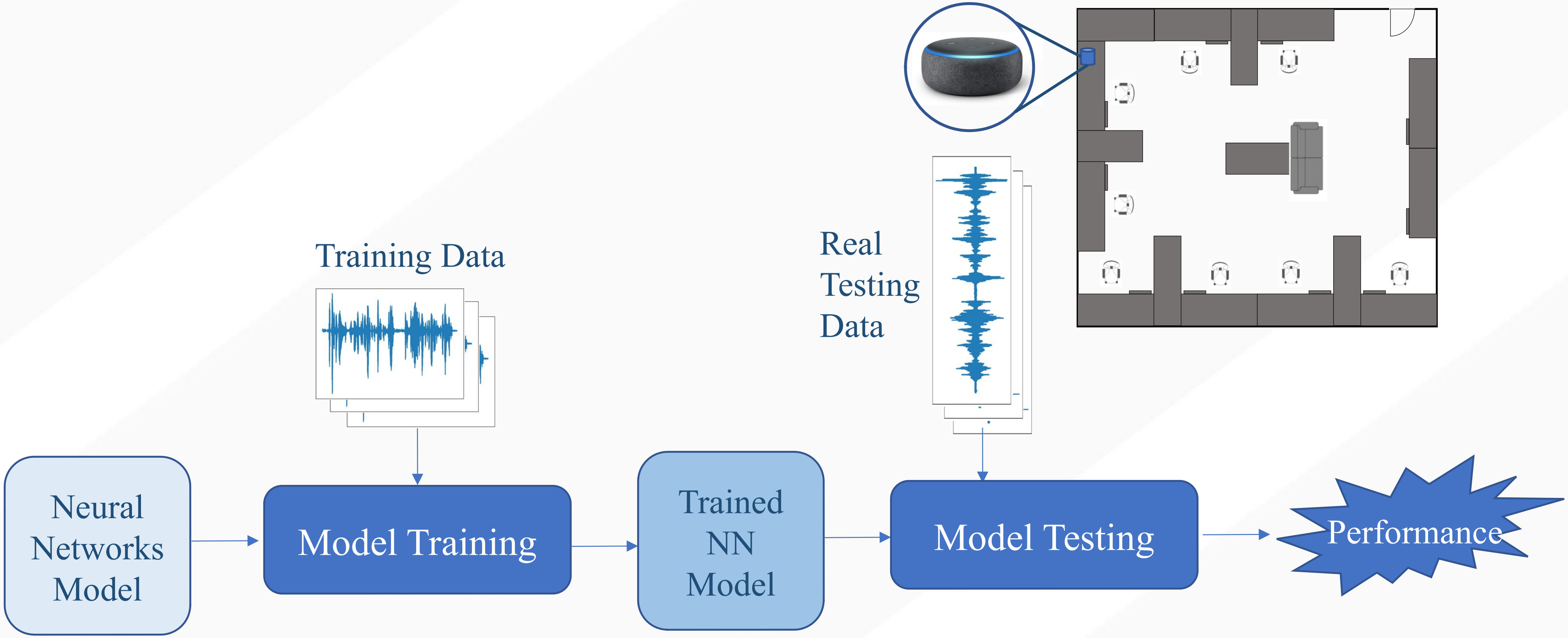
SQEE: A Machine Perception Approach to Sensing Quality Evaluation at the Edge by Uncertainty Quantification

Shuheng Li, Jingbo Shang, Rajesh K. Gupta, Dezhi Hong
UC San Diego, Amazon

Speaker: Shuheng Li
Email: shl060@ucsd.edu

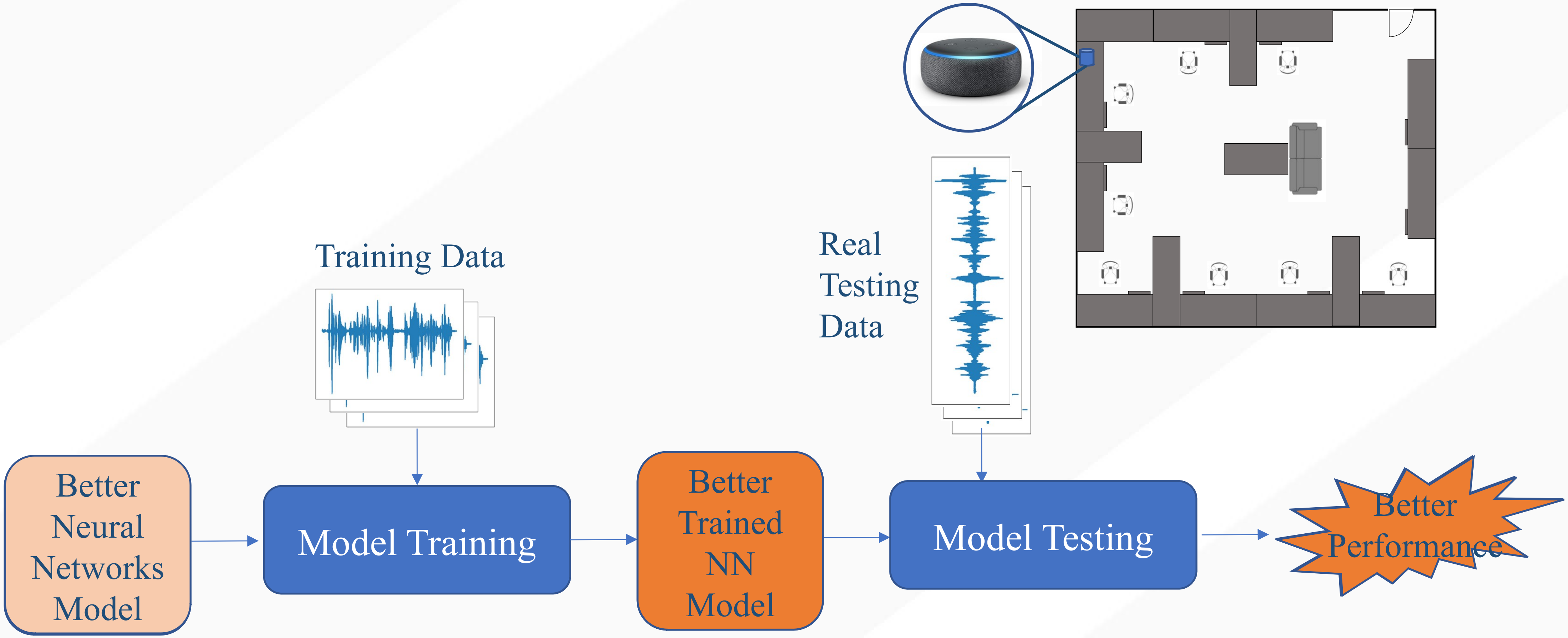
Introduction

- How to improve the performance of a smart sensing system?



Introduction

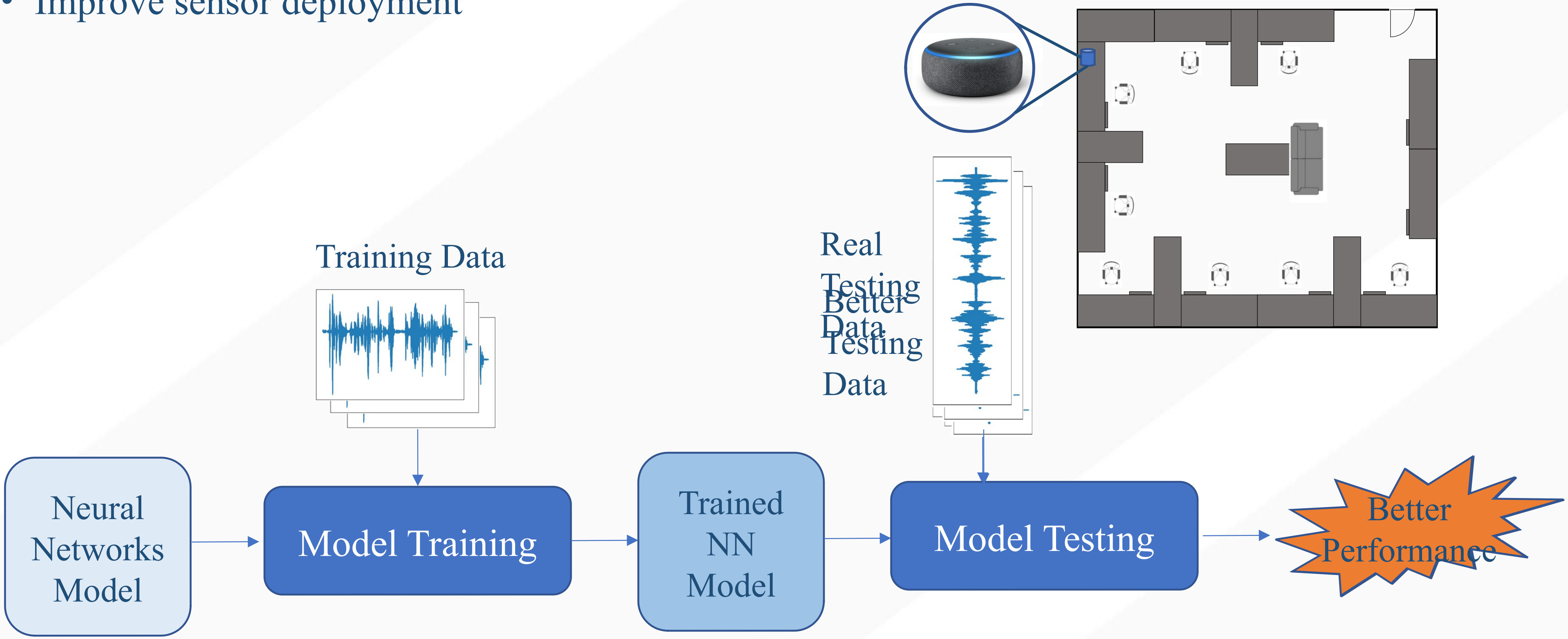
- How to improve the performance of a smart sensing system?
  - Improve the NN model



UC San Diego

Introduction

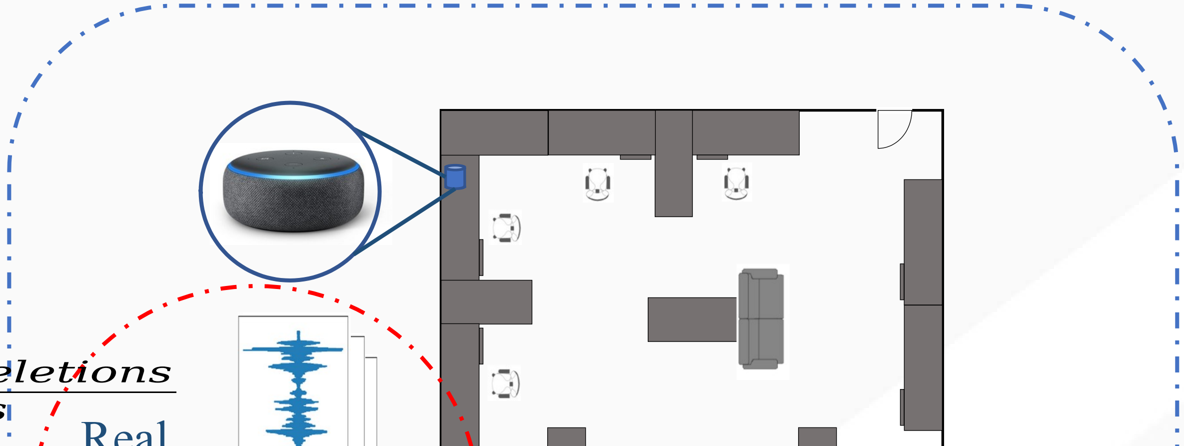
- How to improve the performance of a smart sensing system?
  - Improve the NN model
  - Improve sensor deployment

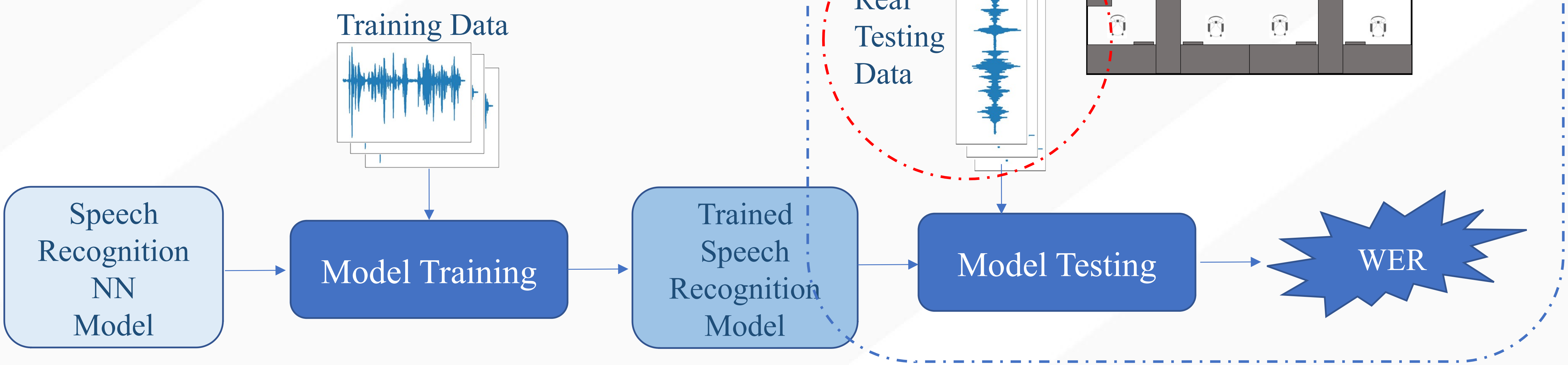


Introduction

- In the context of speech sensing, the goal is speech recognition
  - NN model
    - Input: audio signal
    - Output: recognized natural language sentence
  - Performance
    - Word error rate (WER)
    - WER compares the prediction and the actual spoken words:

$$WER = \frac{Substitutions + Insertions + Deletions}{Total\ number\ of\ words}$$

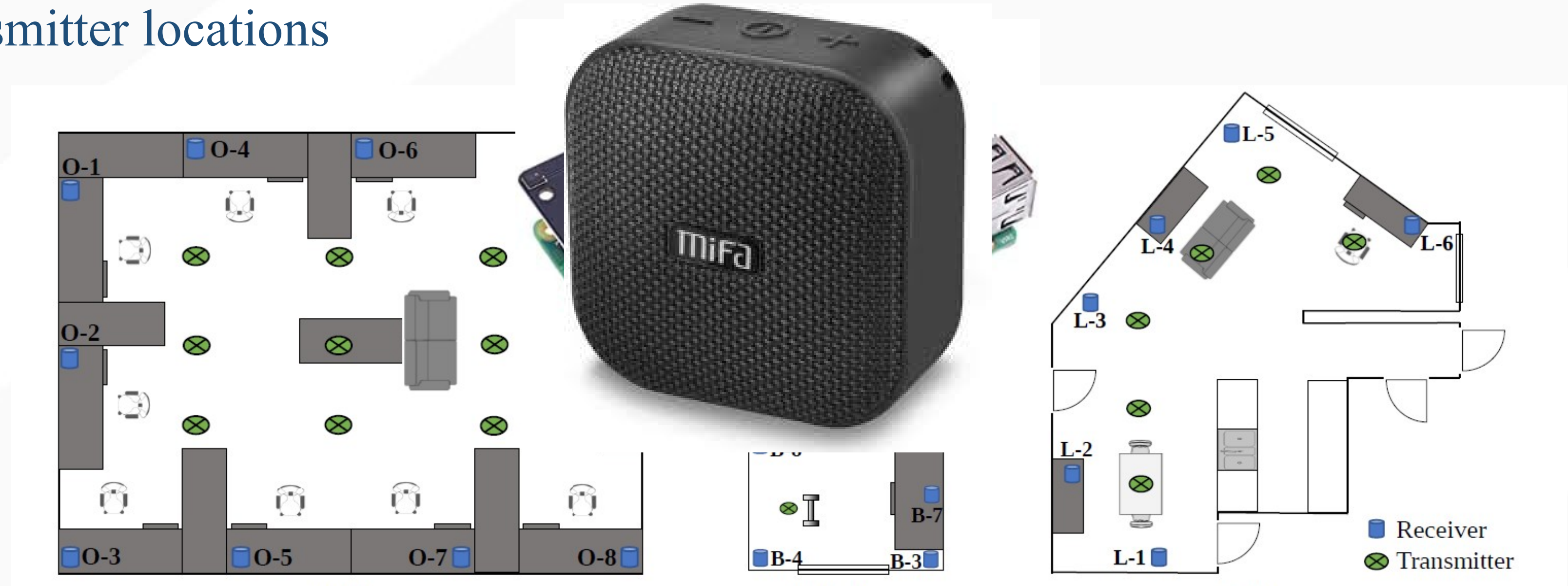




UC San Diego

Benchmark Dataset

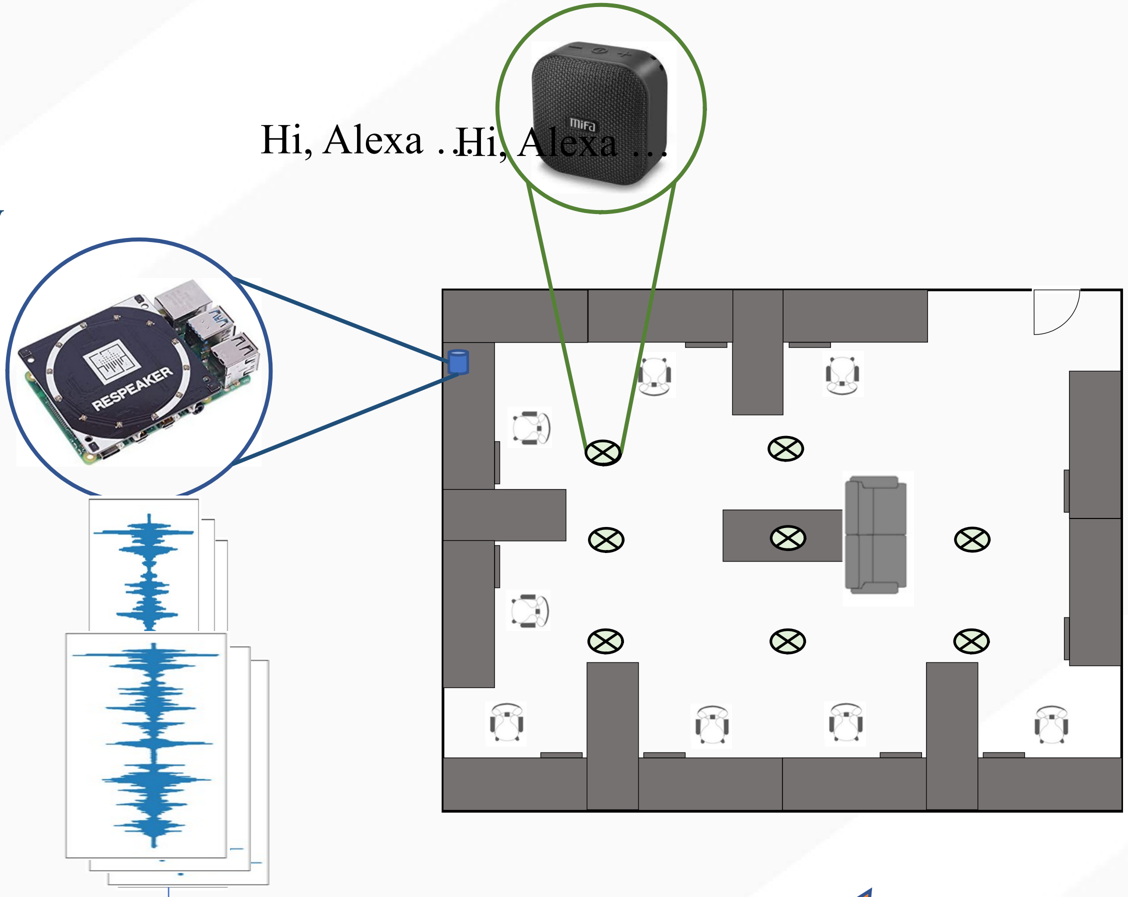
- Receiver: 2 ReSpeaker (4 Mic Array)
- Transmitter: 1 MiFA speaker
- We tested the following three environments:
  - Receiver locations
  - Transmitter locations

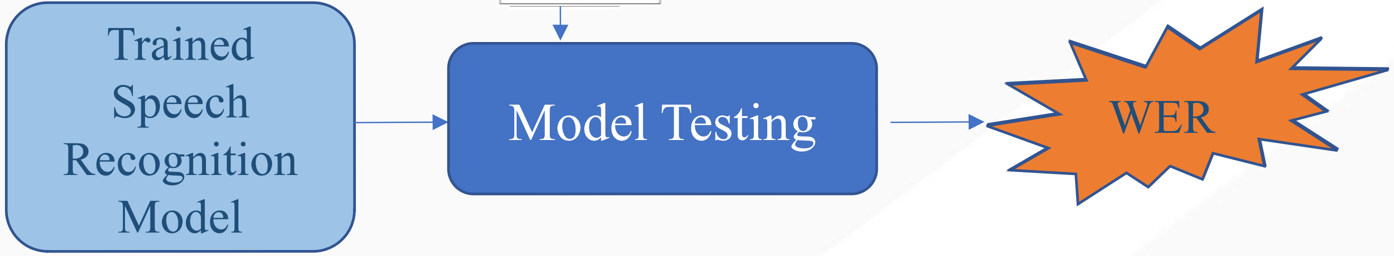


(a) Office (b) Bedroom (c) Living room

Benchmark Dataset

- How to find the best location?
  1. Deploy the receiver(s)
  2. Deploy the transmitter and play a set of N speech examples
  3. Repeat 2 on other transmitter locations
  4. Calculate the WER
  5. Repeat 1 on other receiver locations

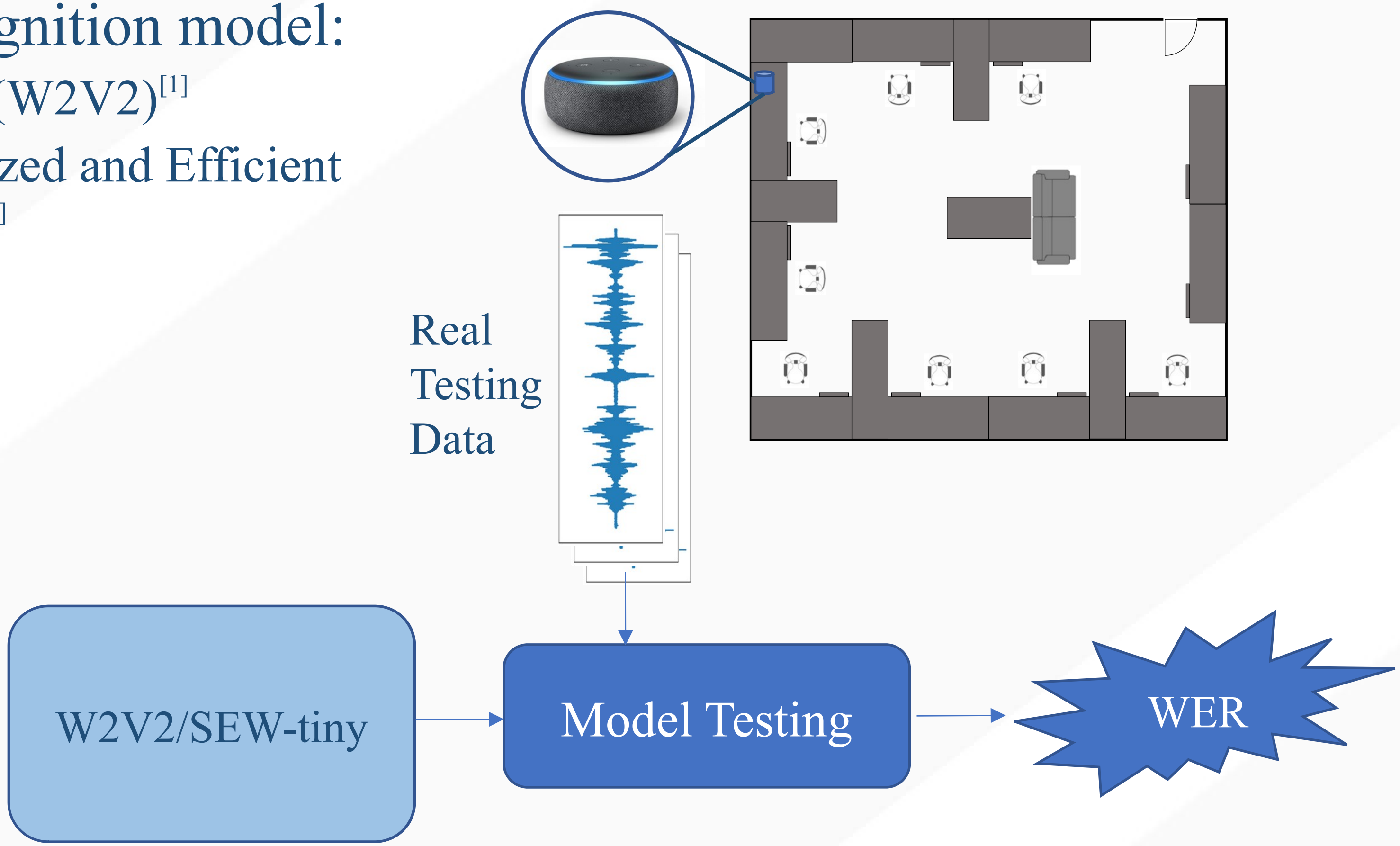




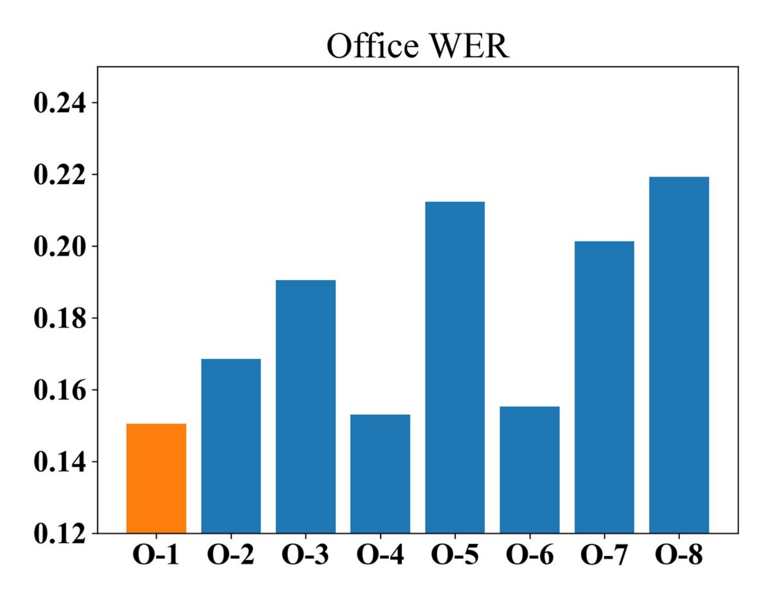
Benchmark Dataset

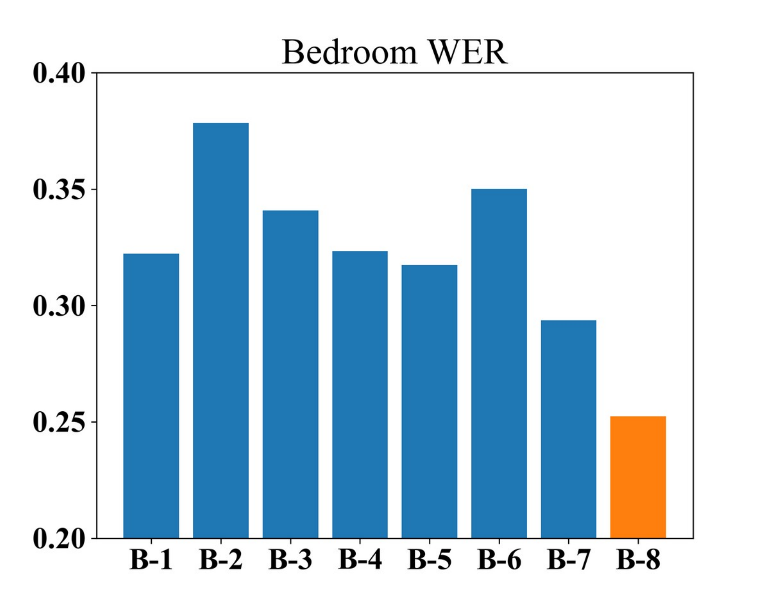
UC San Diego

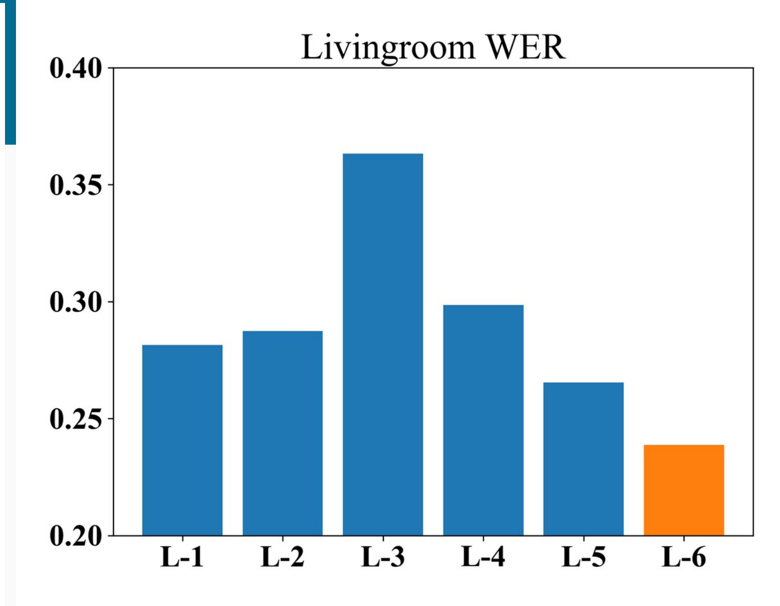
- Trained speech recognition model:
  - Online: Wav2Vec2 (W2V2)[1]
  - Offline: Tiny Squeezed and Efficient W2V2 (SEW-tiny)[2]

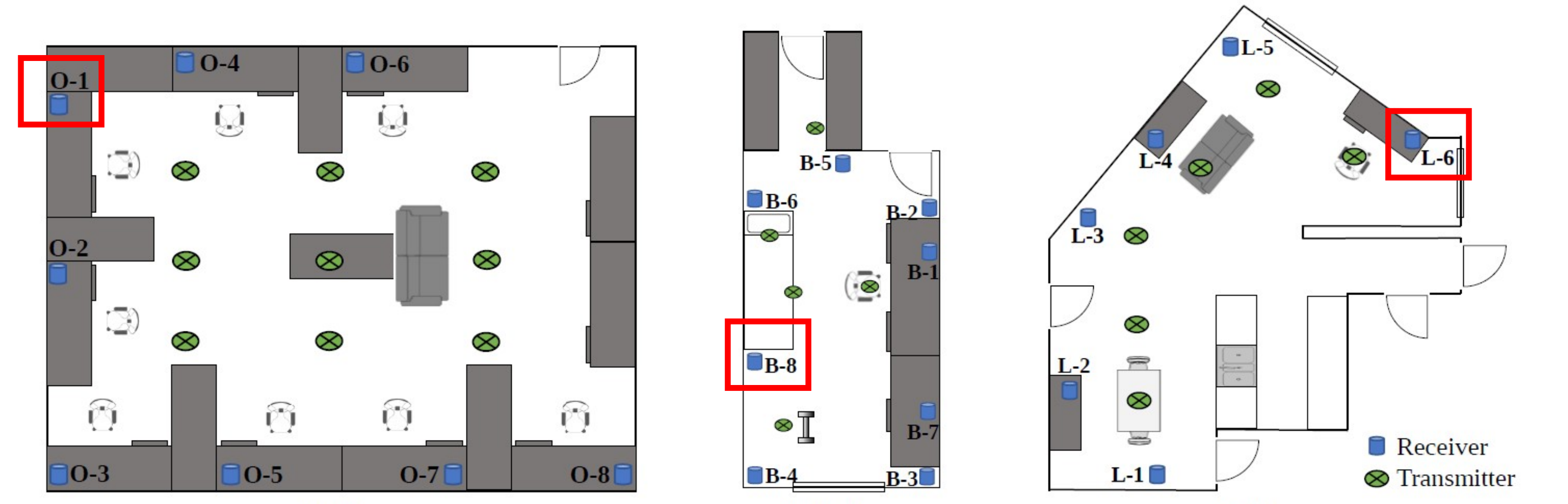


[1] Baevski, Alexei, et al. "Wav2Vec 2.0: A framework for self-supervised learning of speech representations."
[2] Wu, Felix, et al. "Performance-Efficiency Trade-offs in Unsupervised Pre-training for Speech Recognition."



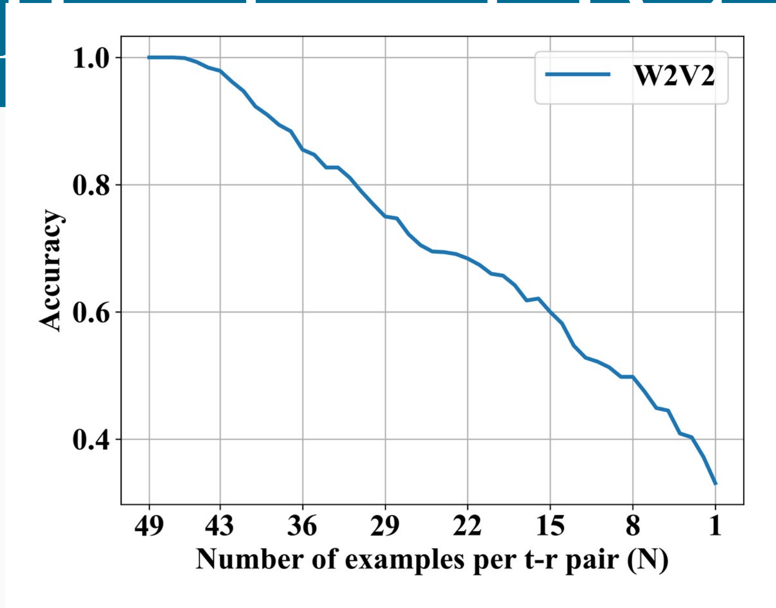




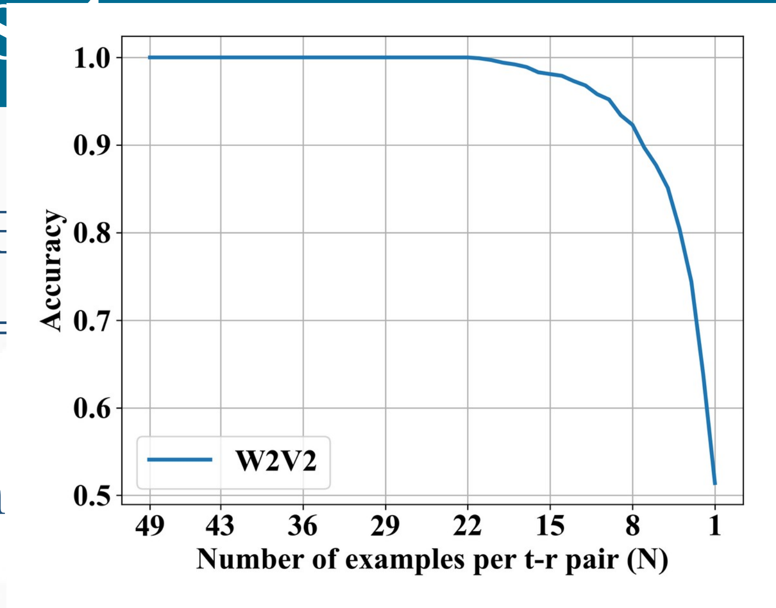


(a) Office (b) Bedroom (c) Living room

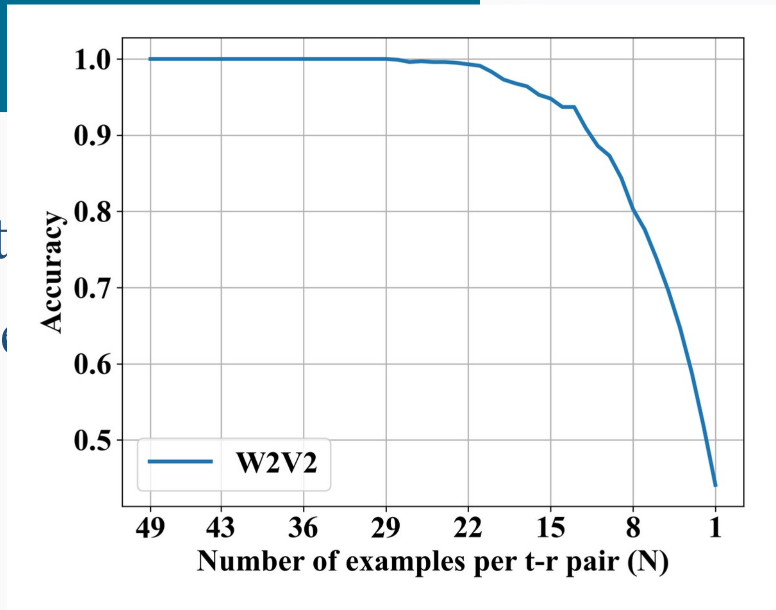
[1] Baevski, Alexei, et al. "Wav2Vec 2.0: A framework for self-supervised learning of speech representations."
[2] Wu, Felix, et al. "Performance-Efficiency Trade-offs in Unsupervised Pre-training for Speech Recognition."



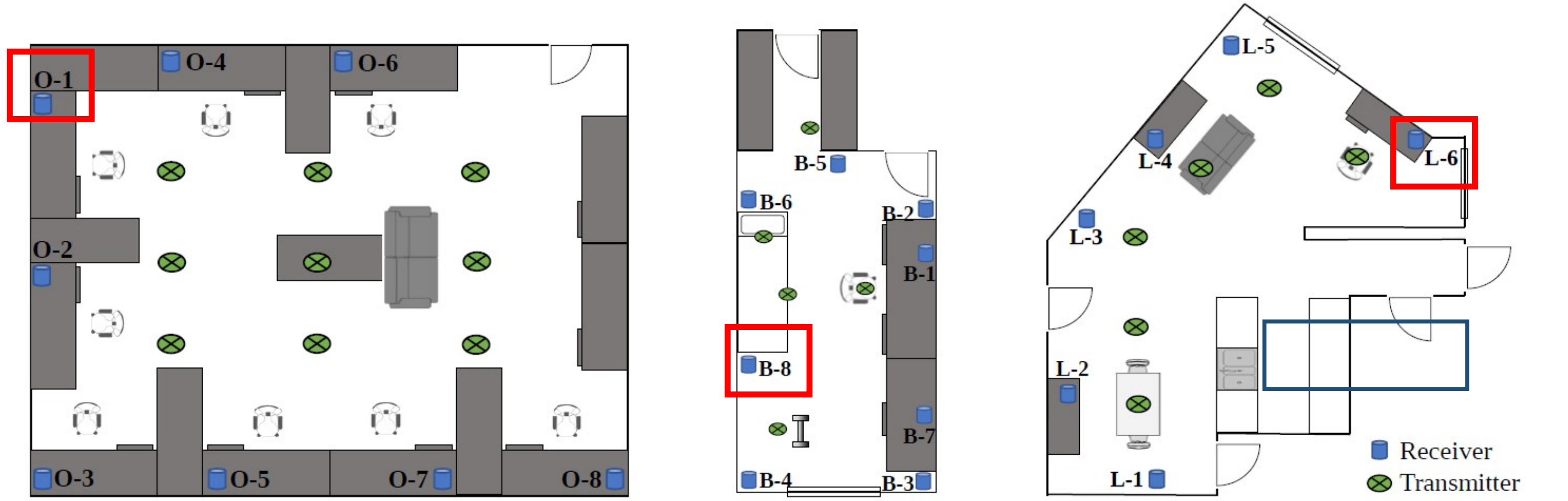
Office



Bedroom



Livingroom

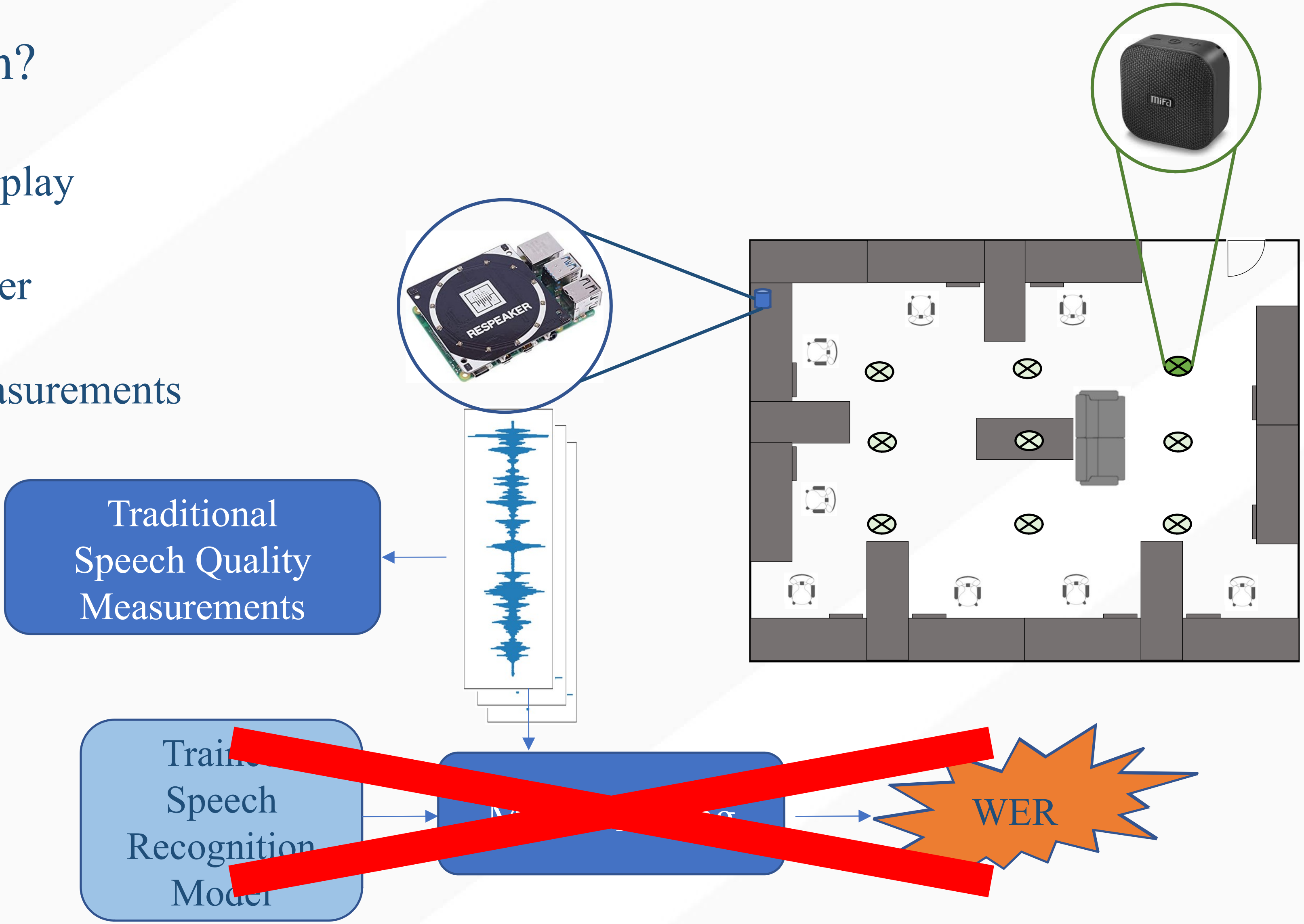


(a) Office (b) Bedroom (c) Living room

Related Works

UC San Diego

- How to find the best location?
  1. Deploy the receiver(s)
  2. Deploy the transmitter and play a set of N speech examples
  3. Repeat 2 on other transmitter locations
  4. Use Traditional quality measurements
     Calculate the WER
  5. Repeat 1 on other receiver locations



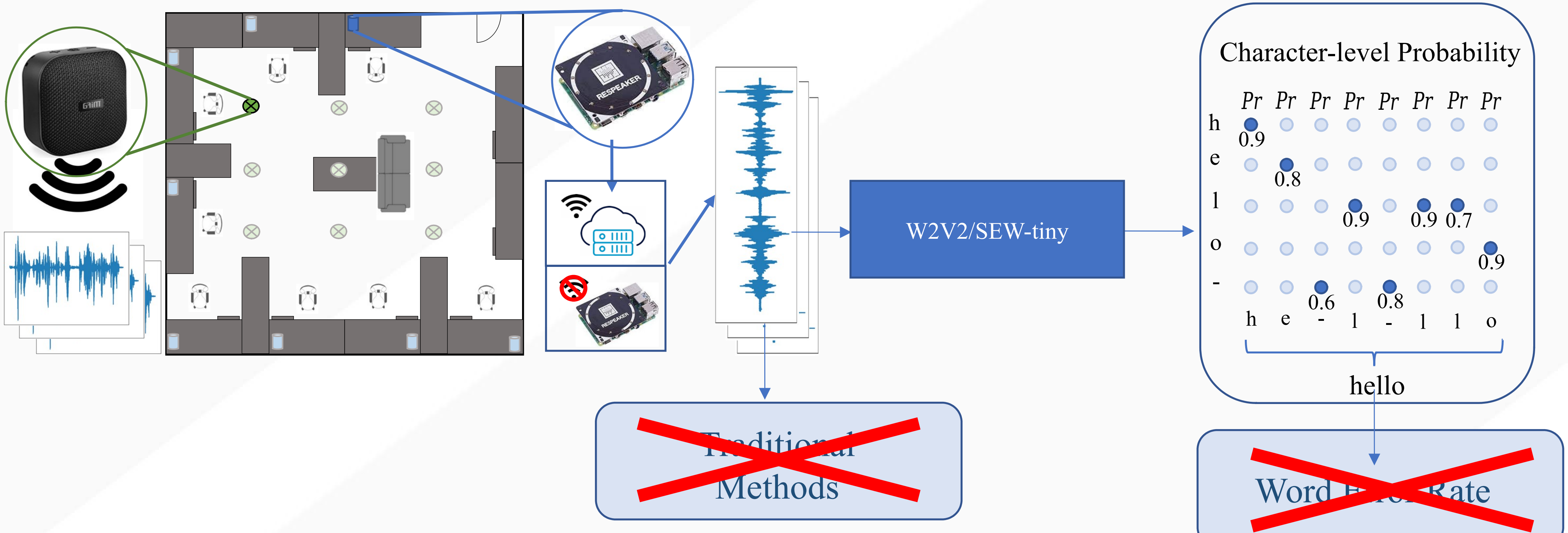
Related works

UC San Diego

- Existing speech sensing quality measurements are ineffective.

Previous Metrics	Office	Bedroom	Living room	Methodology
SNR	✖	✖	◯	Statistical
SSIM	✖	✖	✖	Statistical
PESQ (ITU R.862)	✖	◯	◯	Human Perceptual
DNSMOS	✖	◯	◯	Human Perceptual
STOI	✖	✖	✖	Statistical

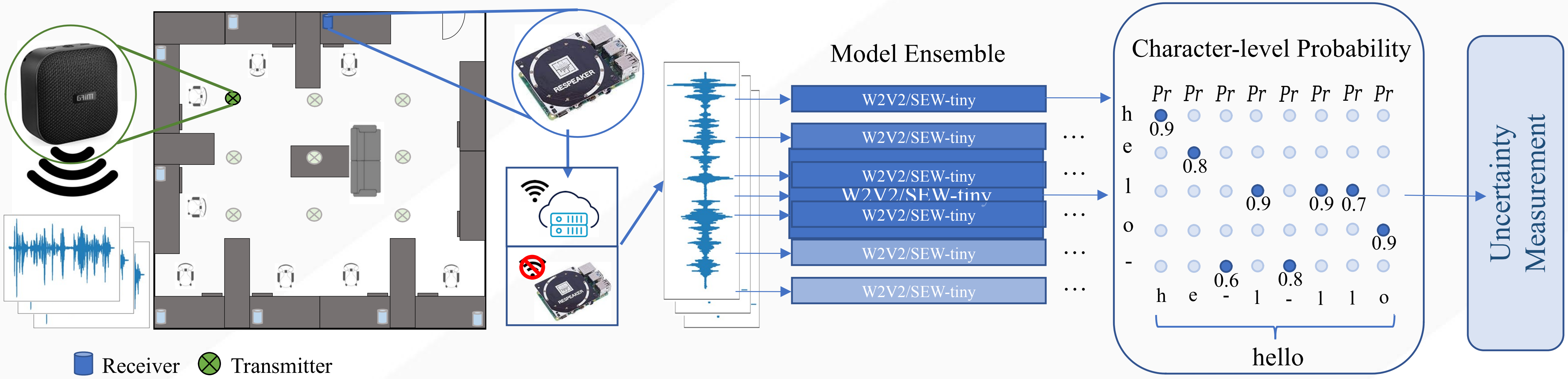
Methodology



- Traditional methods take input the raw speech signal, are at the lowest abstraction level.
  - Informative but noisy
- Speech recognition model predicts the character-level probability. To get the final prediction, the argmax is taken and a merging algorithm is applied. WER is at the highest abstraction level.
  - Information lose

Methodology

UC San Diego



- We estimate the model prediction uncertainty for data collected at each candidate location
  - Model Ensemble
    - Monte-Carlo Dropout [3]
    - All Layer (AL) / Last Layer (LL)
  - Uncertainty Terms
    - *Total uncertainty* (TU) as the entropy of the average over the probability of each model
    - *Data uncertainty* (DU) as the average over the entropy of the probability of each model

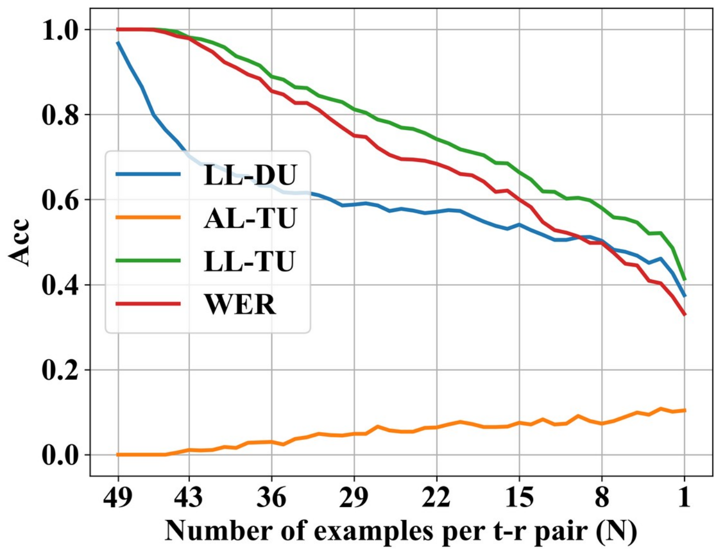
$$\mathcal{H}[y|x] \approx \sum_{y}\left[\frac{1}{M}\sum_{m=1}^{M} -\Pr(y|x,\theta^{(m)})\right] \ln\left[\frac{1}{M}\sum_{m=1}^{M} \Pr(y|x,\theta^{(m)})\right]$$

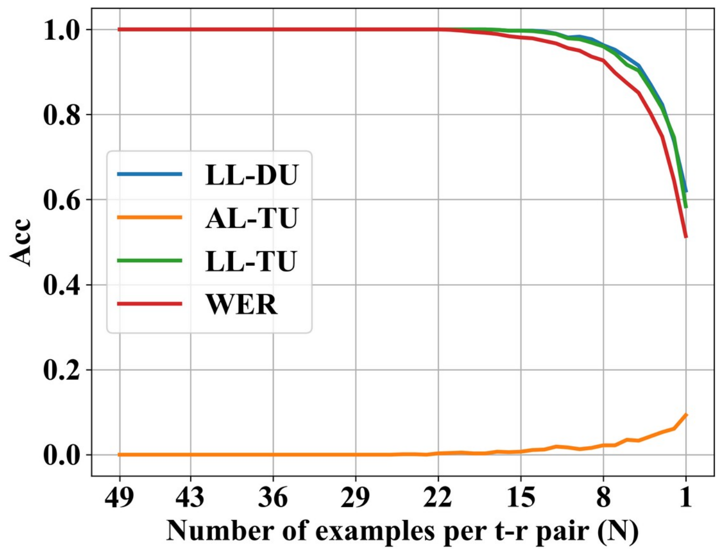
$$\mathbb{E}_{\theta}[\hat{\mathcal{H}}[y|x,\theta]] \approx \frac{1}{M}\sum_{m=1}^{M}\sum_{y} -\Pr(y|x,\theta^{(m)})\ln\Pr(y|x,\theta^{(m)}).$$

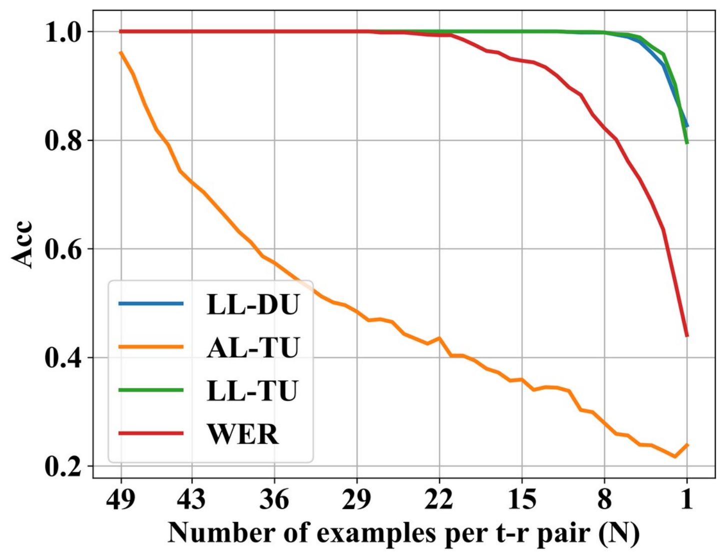
[3] Gal, Yarin, et al. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning."

Evaluation

- LL/AL: Last Layer dropout / All Layer dropout
- TU/DU: Total uncertainty / Data Uncertainty
- The proposed method LL-TU outperforms the compared variants and the WER baseline



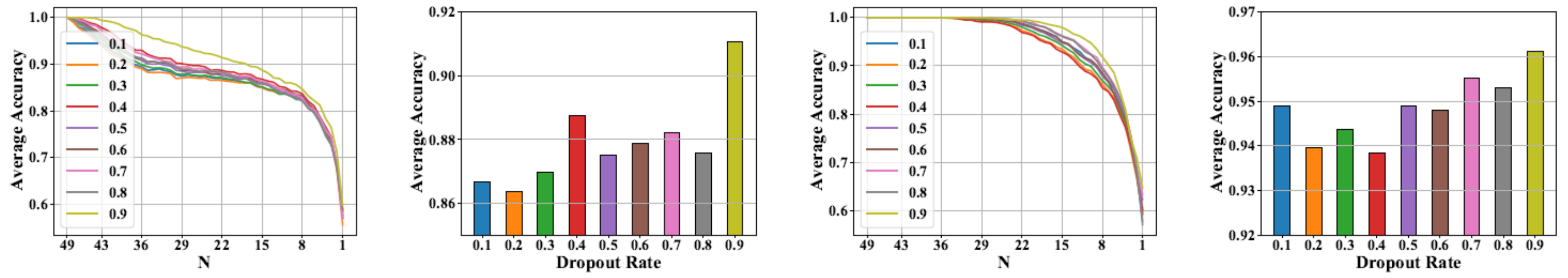




Office Bedroom Livingroom

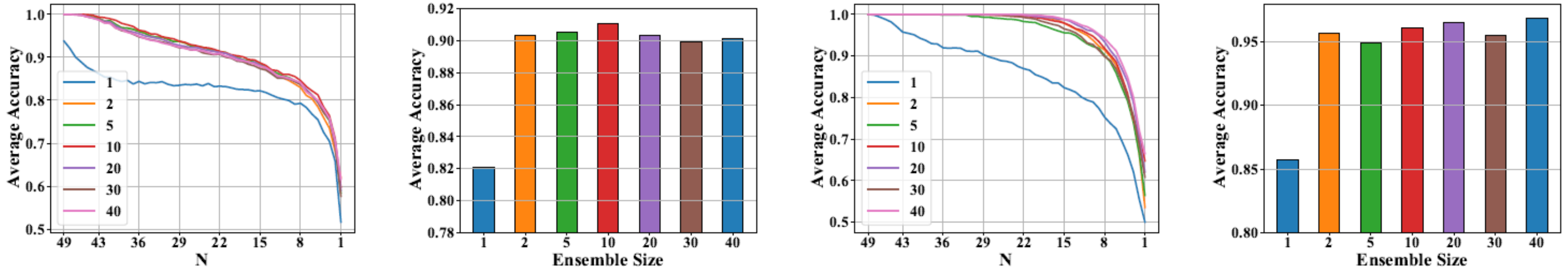
Evaluation – Hyperparameter

- We also study the tuning of hyperparameters for LL-TU:
  - Dropout rate



(a) W2V2 (b) W2V2 (c) SEW-tiny (d) SEW-tiny

  - Ensemble size



(a) W2V2 (b) W2V2 (c) SEW-tiny (d) SEW-tiny

Evaluation – Real World Efficacy

UC San Diego

- SQEE achieves good top-3/top-2 accuracy with only 2 examples needed.

office	Top-1 Acc	Top-3 Acc	T-R time (s)
N = 1	0.414	0.961	15.92
N = 2	0.486	0.972	25.83
N = 3	0.521	0.991	32.27
N = 4	0.520	0.998	45.51
N = 5	0.546	0.998	53.85
N = 10	0.604	1.000	87.69

bedroom	Top-1 Acc	Top-3 Acc	T-R time (s)
N = 1	0.584	0.887	15.52
N = 2	0.746	0.958	26.03
N = 3	0.814	0.982	34.60
N = 4	0.861	0.996	42.62
N = 5	0.903	1.000	55.54
N = 10	0.977	1.000	91.50

living room	Top-1 Acc	Top-2 Acc	T-R time (s)
N = 1	0.796	0.937	17.06
N = 2	0.902	0.983	25.15
N = 3	0.958	0.999	33.15
N = 4	0.972	1.000	46.43
N = 5	0.989	1.000	54.83
N = 10	0.999	1.000	95.13

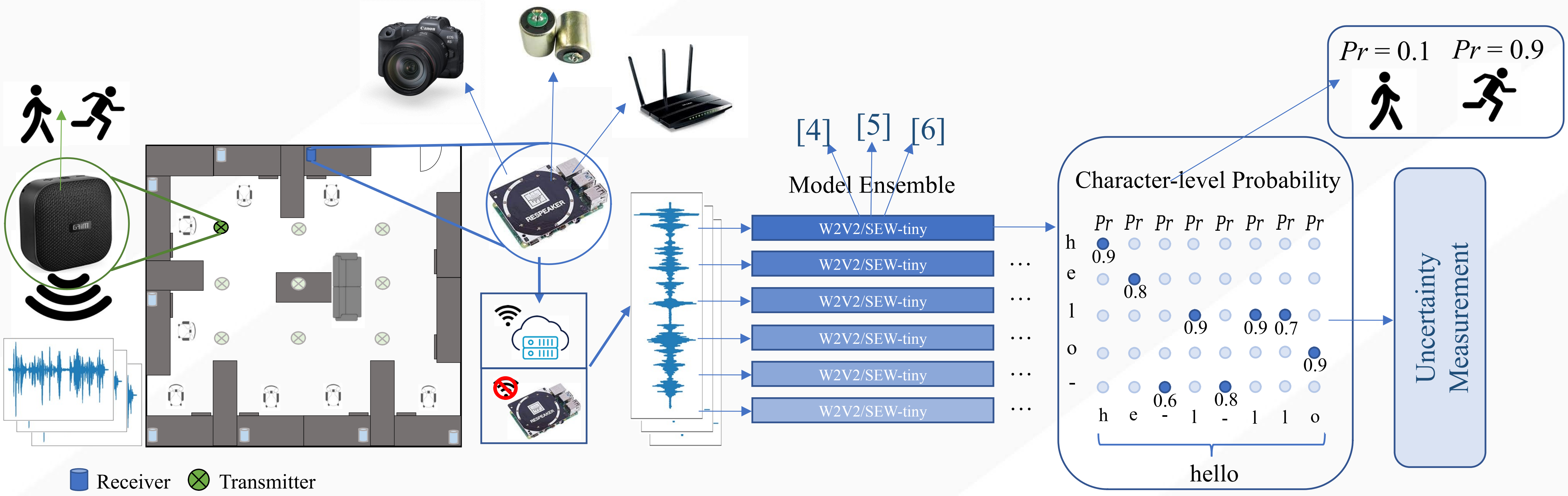
- To outperform the performance of SQEE with N = 2, WER needs (7, 4, 13) examples in the three environments respectively.

WER	Office	Bedroom	Living room
N =	7	4	13

- Our improvement is 2 – 6 times in term of time efficiency.

Discussion – Generalization

- The framework of SQEE can be generalized to other sensing modalities.



[4] Karen, Simonyan, et al. "Very deep convolutional networks for large-scale image recognition."
[5] Asif, Khan, et al. "Structural vibration-based classification and prediction of delamination in smart composite laminates using deep learning neural network."
[6] Guohao, Lan, et al. "MetaSense: Boosting RF sensing accuracy using dynamic metasurface antenna."

Conclusion – Q&A

UC San Diego

- We proposed SQEE, a Machine Perception Approach to Sensing Quality Evaluation at the Edge
  - Improve the performance of NN models from the deployment
  - Online / *Offline settings*
- Built benchmark dataset
  - Three different environments
  - Benchmark evaluation
- *Methodology*
  - *Model ensemble*
  - *Uncertainty quantification for speech recognition*
- Extensive experiments
  - Outperform baseline method
  - Real-world efficacy
- Generalization Discussion