ACOUSTIC FENCING USING MULTI-MICROPHONE SPEAKER SEPARATION (6090)

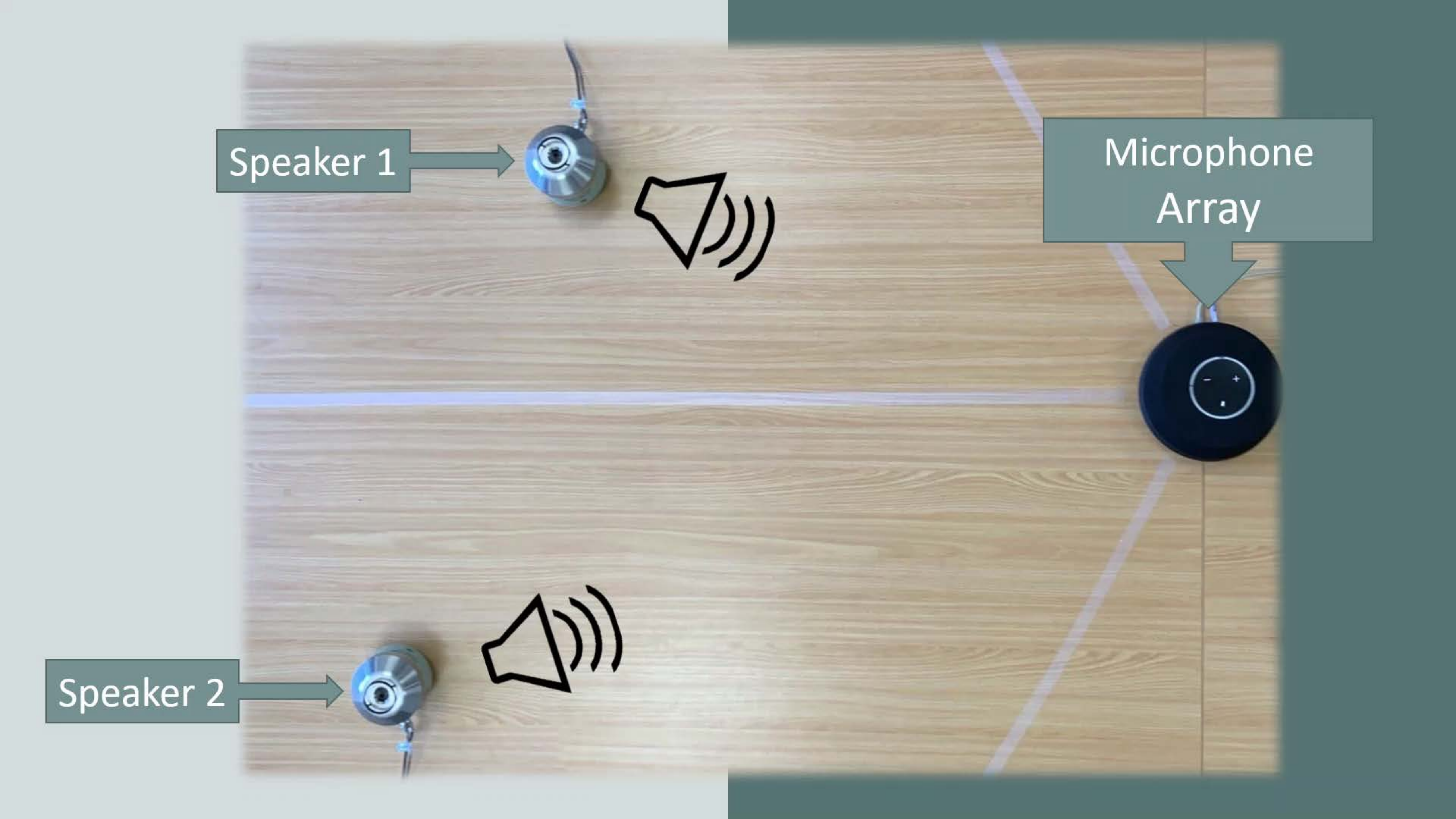
Students: Orel Ben-Reuven and Tomer Fait
Instructor: Amir Ivry

August 2021

The Andrew and Erna Viterbi Faculty of ELECTRICAL & COMPUTER ENGINEERING

TECHNION Israel Institute of Technology

SIPL Signal and Image Processing Lab.


Speaker 1
Speaker 2
Microphone Array

OUTLINE

- Challenge and motivation
- Algorithm overview
- Defining evaluation criteria
- Simulated recordings
- Real recordings

THE CHALLENGE

- Acoustic fencing
  - Attenuate speakers outside reception area
  - Do not distort speakers inside reception area
- Speakers outside reception area may be dominant
- Signals are noisy and reverberant

SIGNIFICANCE AND MOTIVATION

- Diarization
- Identify locations of speakers
- Prioritizing speakers according to location
- Highlight specific speaker among cluttered environment
- Automatic transcription and translation

PROBLEM FORMULATION

- M mics and N speech sources in noisy reverberant enclosure

- i_{th} speech signal $s^i(n)$ captured by m_{th} microphone:

$$z_m(n) = \sum_{i=1}^{N} \left(s^i(n) * h_m^i(n)\right) + W_m(n)\,,$$

where $h_m^i(n)$ is the RIR relating the i_{th} speaker to the m_{th} microphone and $W_m(n)$ is the additive AWGN

PROBLEM FORMULATION – CONT.

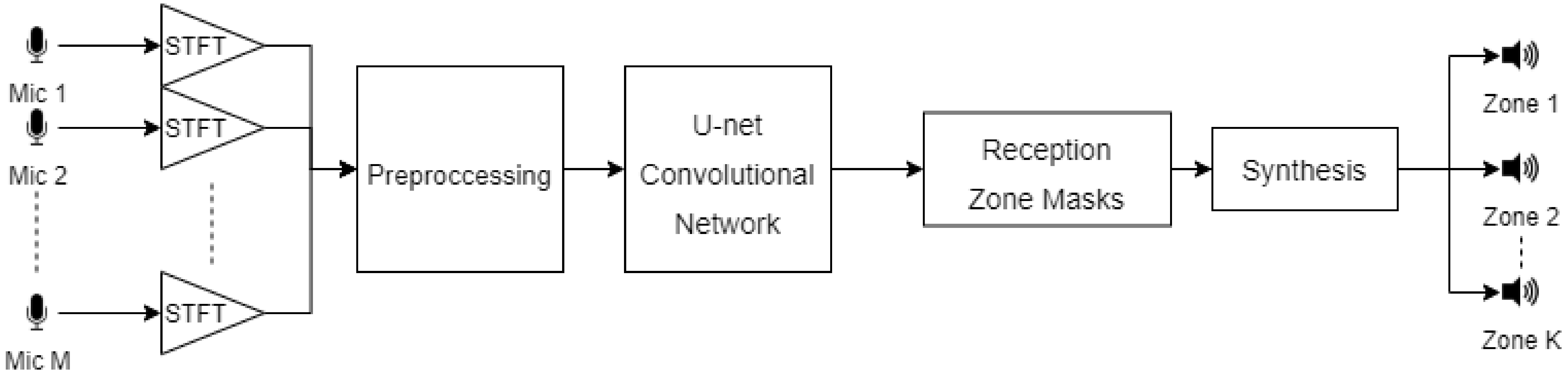
- Let θ_1, θ_2 be two reception zones
- Let $y_1(n), y_2(n)$ be the output signals of the system
- The goal – include only speakers located inside θ_k in $y_k(n)$

DDESS

- Deep direction estimation for speech separation (DDESS) [1] separates speakers in the STFT domain

- DDESS employs:
  - U-net network [2]
  - Classification of each TF bin to DOA
  - Reconstruction by multiplying U-net masks with reference microphone

SCHEME OF DDESS SOLUTION



DDESS CONT.

- 9 microphones (8 + ref.), 2 reception zones, and 2 speakers
- Each microphone contributes a TF image with L time frames and K frequency bins - Z_i
- Calculate the phase difference between each of the 8 microphones and the ref:

$$\angle \frac{Z_i}{Z_{ref}}$$

- The inputs to the net are the sine and cosine of these differences.

DDESS CONT.

- Input - 16 TF images with L time frames and K frequency bins - R
- For each pair $\{(l,k) \mid l = 1, \dots, L\ ; k = 1, \dots, K\}$ we define:

$$r_i^1(l,k) = \cos\left(\angle \frac{Z_i(l,k)}{Z_{ref}(l,k)}\right)$$
$$r_i^2(l,k) = \sin\left(\angle \frac{Z_i(l,k)}{Z_{ref}(l,k)}\right),$$

where Z_i is the STFT transformation of z_i for $i = 2, .., 9$

DDESS CONT.

- Output – for each reception area $\theta_{i\in 1,2}$, we define the mask:

$$\widehat{M}_i(l,k) = p_{l,k}(\theta_i) \, ,$$

where $p_{l,k}(\theta_i)$ is the probability of bin $\{(l,k) \mid l = 1, \ldots, L \, ; k = 1, \ldots, K\}$ to be in reception area $\theta_{i\in 1,2}$

- The estimated reconstructed signals are given by:

$$y_i = iSTFT\{Z_{ref} \cdot \widehat{M}_i\} \mid i = 1,2$$

DDESS CONT.

- Label is given per TF bin, according to dominant speakers' location
- Loss function was the cross-entropy loss
  - Assumes classes have a Bernoulli probability distribution

$$loss(x, class) = -\,log\left(\frac{exp(x[class])}{\sum_j exp(x[j])}\right)$$

$$loss = \sum_{i=1}^{N} loss(i, class[i])$$

EVALUATIONS CRITERIA

- We want to separately evaluate the levels of:
  - Desired signal distortion
  - Interference suppression

- We applied the output mask to 3 STFT signals:
  - Original input signal (mix of both speakers), $r(n)$
  - Desired signal, $p(n)$
  - Interference signal, $b(n)$

NOTATION

$$\widehat{M}$$
$$r(n) \rightarrow \; r'(n)$$
$$p(n) \rightarrow \; p'(n)$$
$$b(n) \rightarrow \; b'(n)$$

From the linearity of STFT we get that:

$$p(n) + b(n) = r(n)$$

$$\Longrightarrow p'\,(n) + b'(n) = r'(n)$$

EVALUATION CRITERIA – CONT.

- Frequency weighted SNR (fwSNRseg) [3] between the clean signal $p(n)$ and $p(n) - p'(n)$
- Output SIR:

$$oSIR = \frac{RMS(p')}{RMS(b')}\ [dB]$$

- SIR gain - the difference between output and input SIRs:

$$iSIR = \frac{RMS(p)}{RMS(b)}\ [dB]$$

$$SIR\ Gain = oSIR - iSIR$$

DATA CORPUS

- TIMIT contains (in English):
  - 6300 sentences
  - 630 speakers
  - 8 major dialect
  - 2000+ textually different sentences
- The test set is 27% (about 40 minutes) of the data set
- Popular database for benchmarking

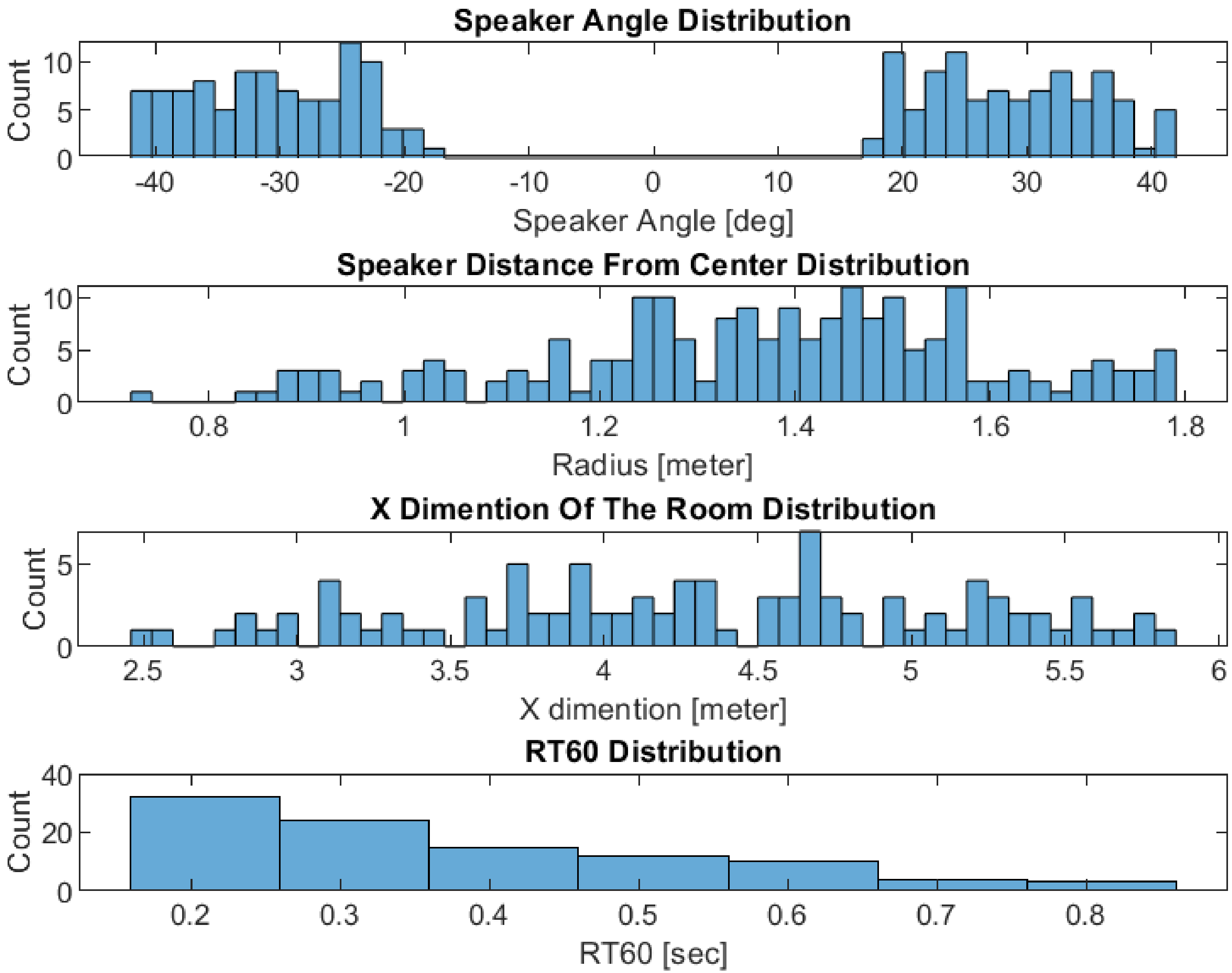
SIMULATION SETUP

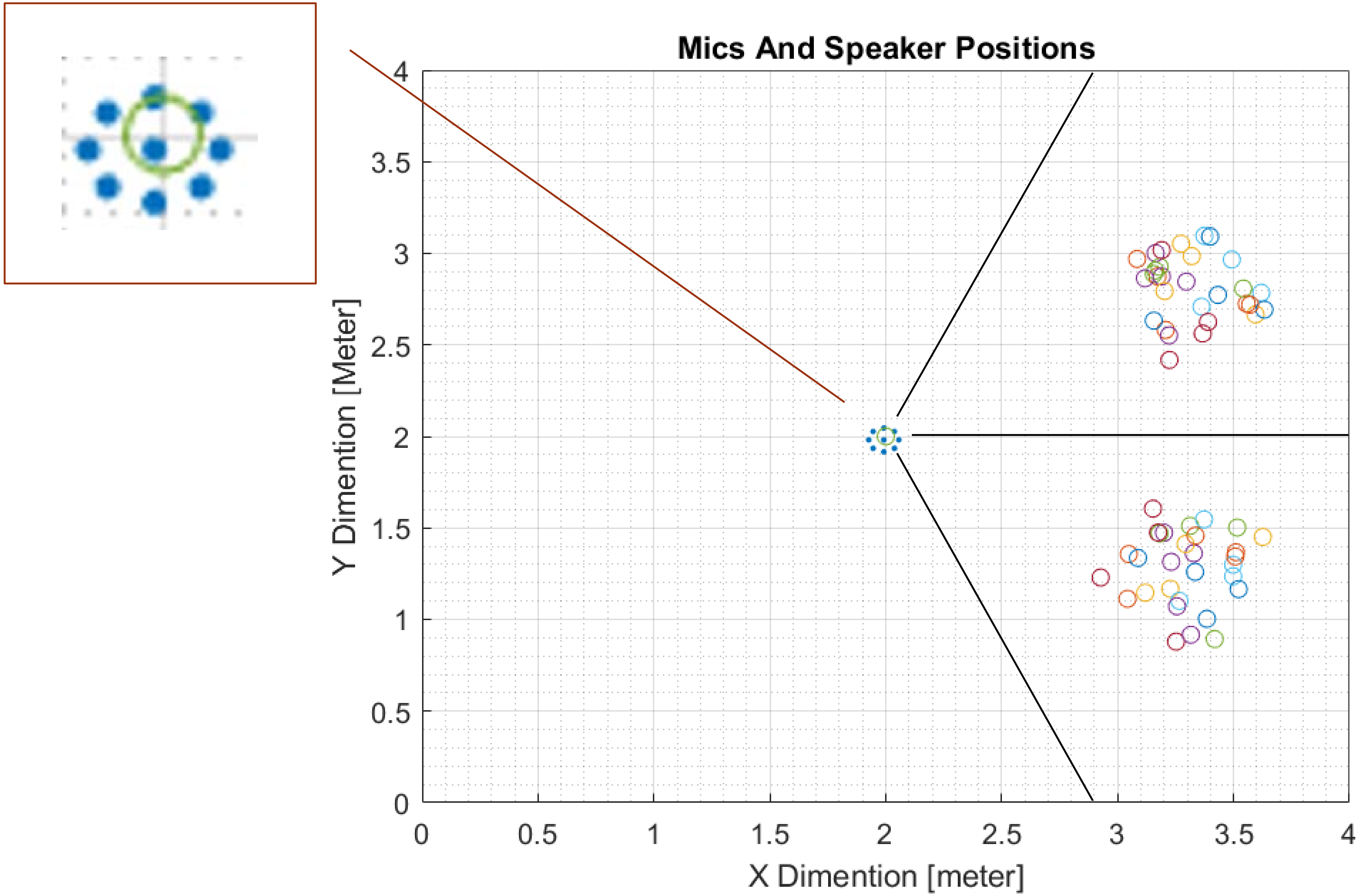
- Simulations include varying room dimensions, speaker/receiver setups, and reverberation times
- Each scenario is randomized from the following distributions:
  - Room dimensions: $(X, Y, Z) \sim N(4.25, 1)$
  - Mics array location: Uniform $[{}^{X}/_{2} \pm 0.05, {}^{Y}/_{2} \pm 0.05, 0.75 \pm 0.05]$
  - Speaker's radius: Uniform $[1.5 \pm 0.3]$ (up to room boundary)
  - Speaker's angle: Uniform $[18° - 42°]$

SIMULATION SETUP

- RIR is generated using image method [4] with $RT_{60} \sim Poisson(0.3)$, bounded by $[0.2, 0.8]$ seconds

- AWGN is added to each microphone with $SNR \sim N(35, 25)$ [dB]


Speaker Angle Distribution
Count
Speaker Angle [deg]
Speaker Distance From Center Distribution
Count
Radius [meter]
X Dimention Of The Room Distribution
Count
X dimention [meter]
RT60 Distribution
Count
RT60 [sec]


Mics And Speaker Positions
Y Dimention [Meter]
X Dimention [meter]
0
0.5
1
1.5
2
2.5
3
3.5
4

SIMULATION RESULTS

Training on 2h TIMIT dataset, the test set results are:

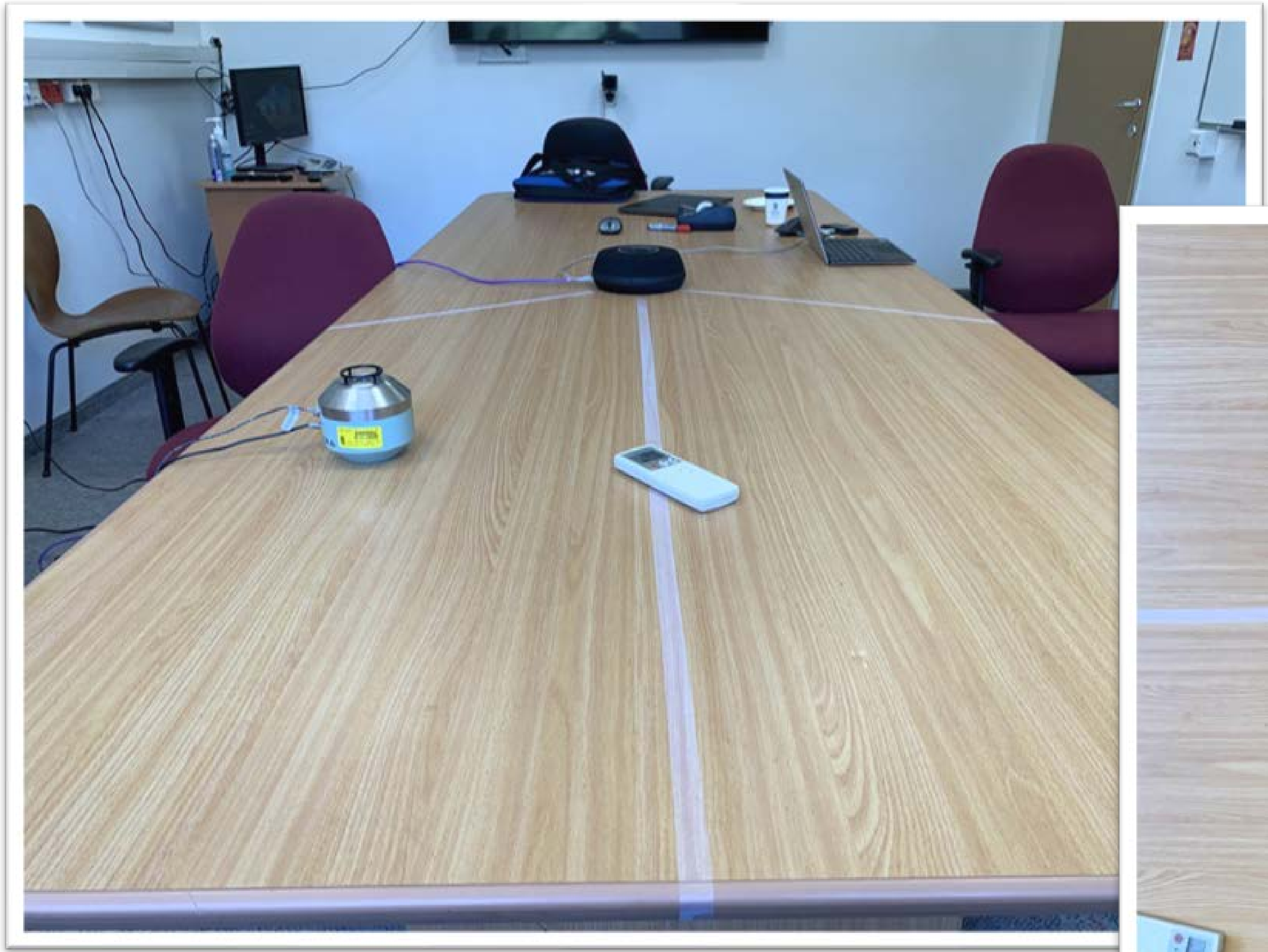
(averaged across sentences)

Evaluation Criterion	Speaker 1	Speaker 2
fwSNR [dB]	13 ± 7	14 ± 7
Output SIR [dB]	12 ± 4	12 ± 5
SIR gain [dB]	12 ± 5	12 ± 5

SIMULATION RESULTS - INSIGHTS

- Speakers well separated (quantitatively and qualitatively)
- Noise is classified into specific zone
- One can try to filter noise by creating a virtual zone
- Time for real recordings ..

REAL RECORDINGS - TRAIN



Fun day at the office ☺

REAL RECORDINGS - TRAIN

- $SIR \sim N(0,16)$[dB]
- AWGN is added to each microphone with $SNR \sim N(35, 25)$ [dB]

Computer's center

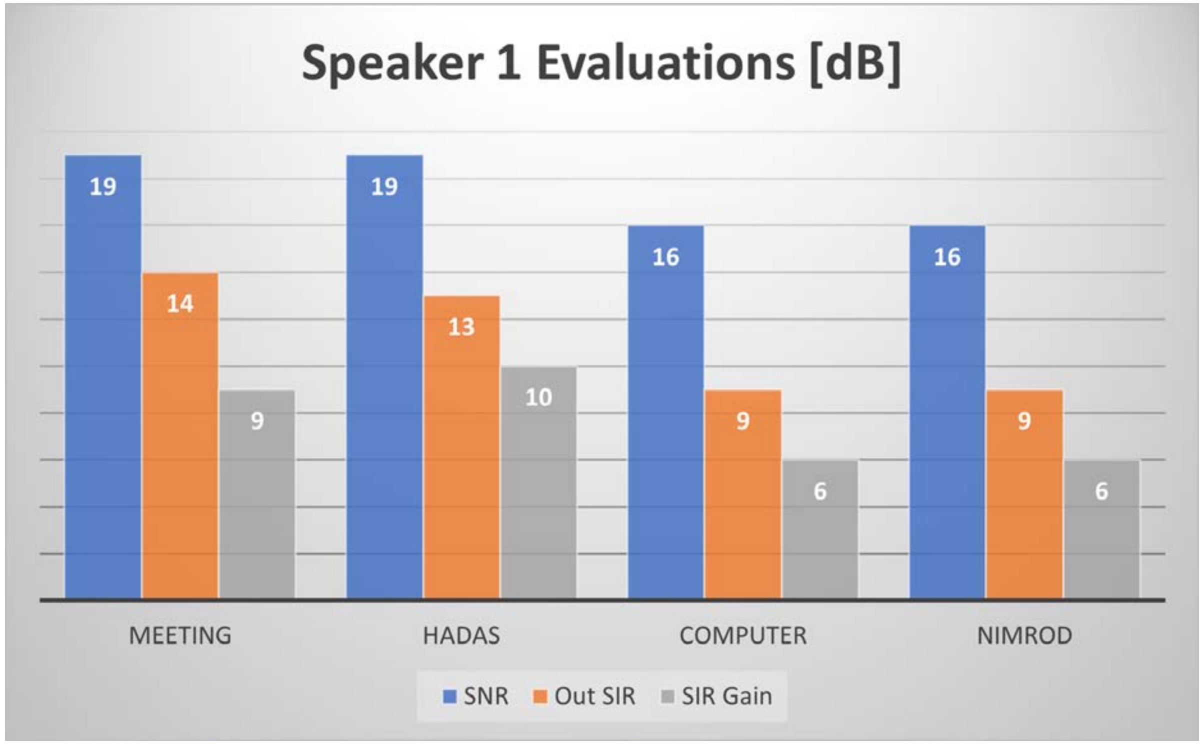
Nimrod

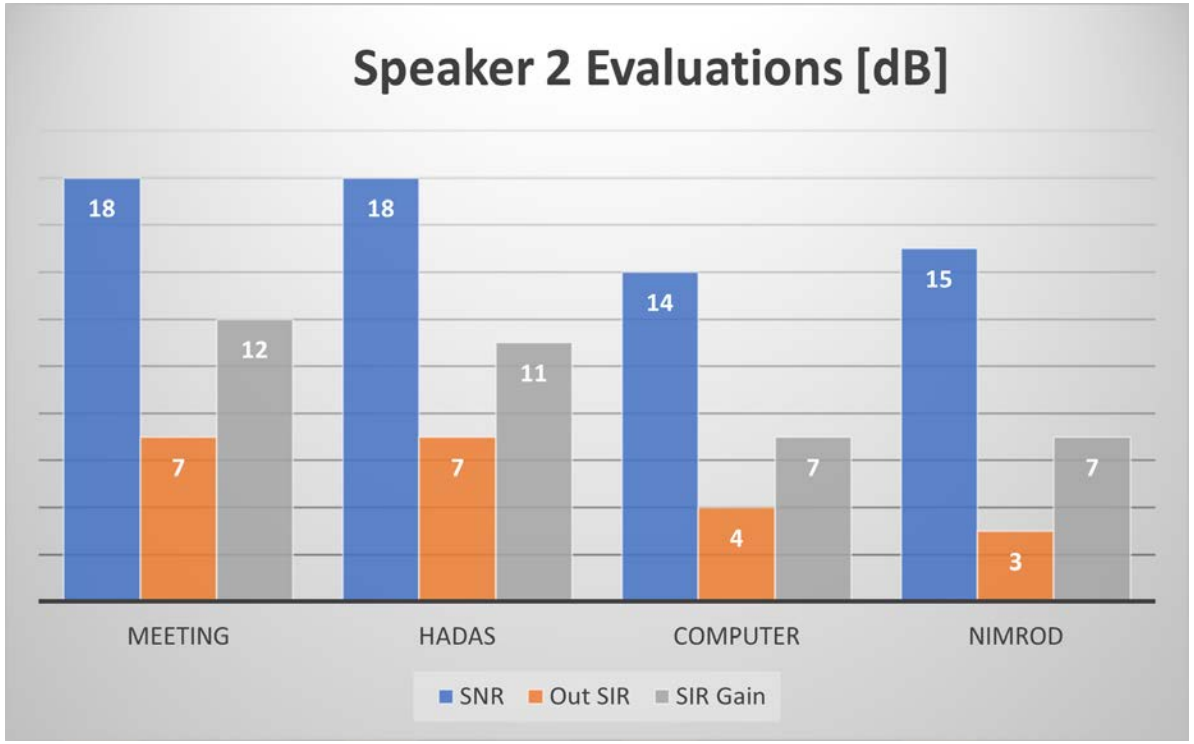
Hadas

REAL RECORDINGS - TEST

- We tested the trained network on 3 additional rooms
- The rooms were not present in the training set
- SIR is randomized, but no AWGN is added

REAL RECORDINGS - RESULTS





REAL RECORDINGS - RESULTS

Speaker	Evaluation criterion	Meetings room	Hadas	Computers' room	Nimrod
1	fwSNR [dB]	19 ± 4	19 ± 4	16 ± 3	16 ± 3
	Output SIR [dB]	14 ± 3	13 ± 4	9 ± 3	9 ± 4
	SIR gain [dB]	9 ± 2	10 ± 3	6 ± 2	6 ± 3
2	fwSNR [dB]	18 ± 4	18 ± 4	14 ± 4	15 ± 4
	Output SIR [dB]	7 ± 2	7 ± 3	4 ± 3	3 ± 3
	SIR gain [dB]	12 ± 3	11 ± 3	7 ± 2	7 ± 2

REAL RECORDINGS - INSIGHTS

Separation is best at meetings room and Hadas' room

High noise in Computers room may impede separation

Obstacles in Nimrod's room may impede separation

REAL RECORDINGS - CONCLUSION

Successful separation of real recordings

Generalization for untrained rooms

Evaluation criteria for Acoustic Fencing

FURTHER INFORMATION

- Our documented code is available at: https://github.com/Orelbenr/acoustic-fencing

- Full report and demo files are also available in the GitHub repository

THANK YOU

QUESTIONS?

References

[1] Chazan, Shlomo E., et al. "Multi-Microphone Speaker Separation based on Deep DOA Estimation." *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019.

[2] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

[3] Philipos C Loizou. *Speech enhancement: theory and practice*. 2nd ed. Boca Ratton: CRC press, 2013.

[4] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” Journal of the Acoustical Society of America , vol. 65, pp. 943–950, Apr. 1979.