# Wavesdropper: Through-wall Word Detection of Human Speech via Commercial mmWave Devices

**Chao Wang**, Feng Lin, Zhongjie Ba,

Fan Zhang, Wenyao Xu, Kui Ren

浙江大學 ZHEJIANG UNIVERSITY

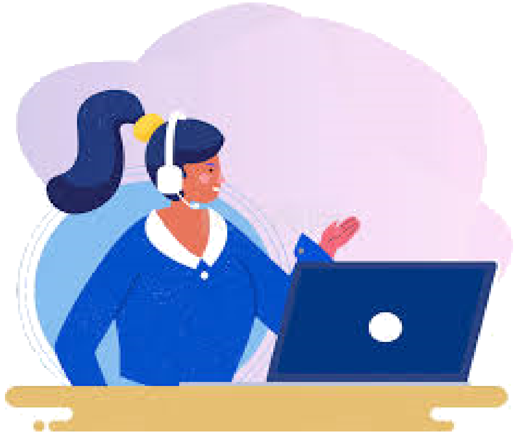
University at Buffalo

# Outline

- Background
- Related Work
- Attack Scenario
- Feasibility Study & Challenge
- System Design & Evaluation
- Countermeasure & Conclusion

# Background

Face-to-face conversation



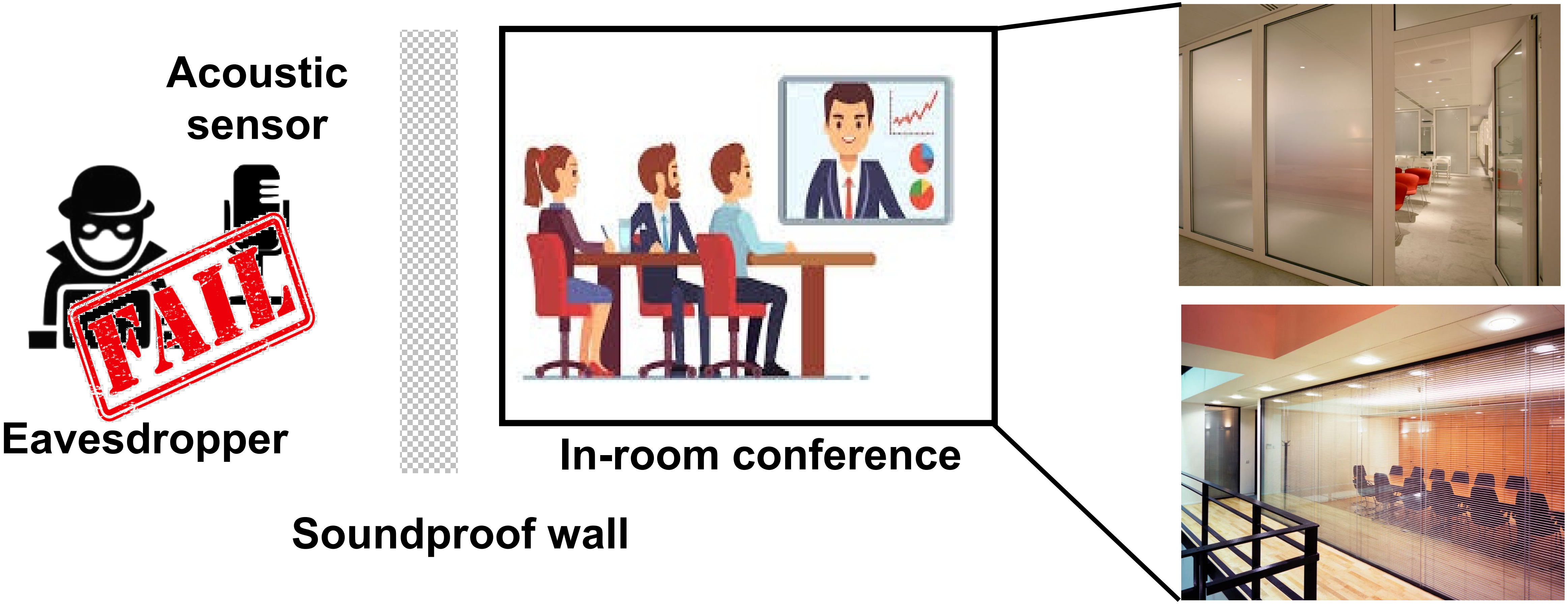
Video call

Enterprise meeting

Virtual conference

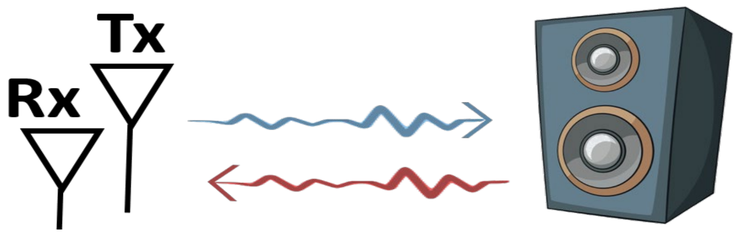
# Background

- Sound isolation is favored to avoid speech leakage.

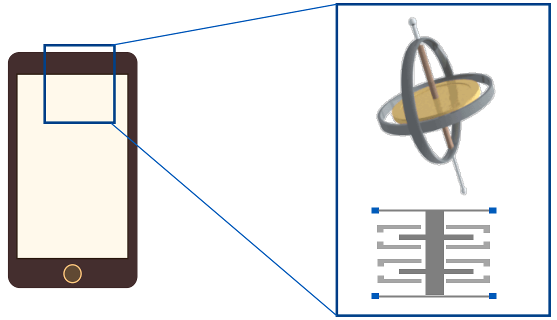


# Related work

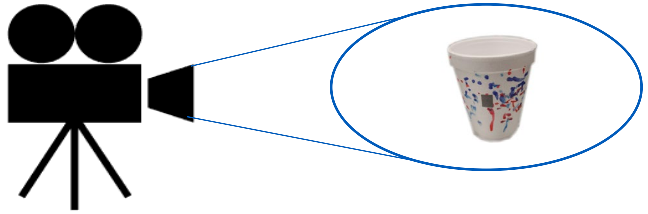
- Vibration-based eavesdropping
  - E.g., RF signals, motion sensors, video cameras, lidars…




RF signals (SenSys'20)



Motion sensors (NDSS'20)

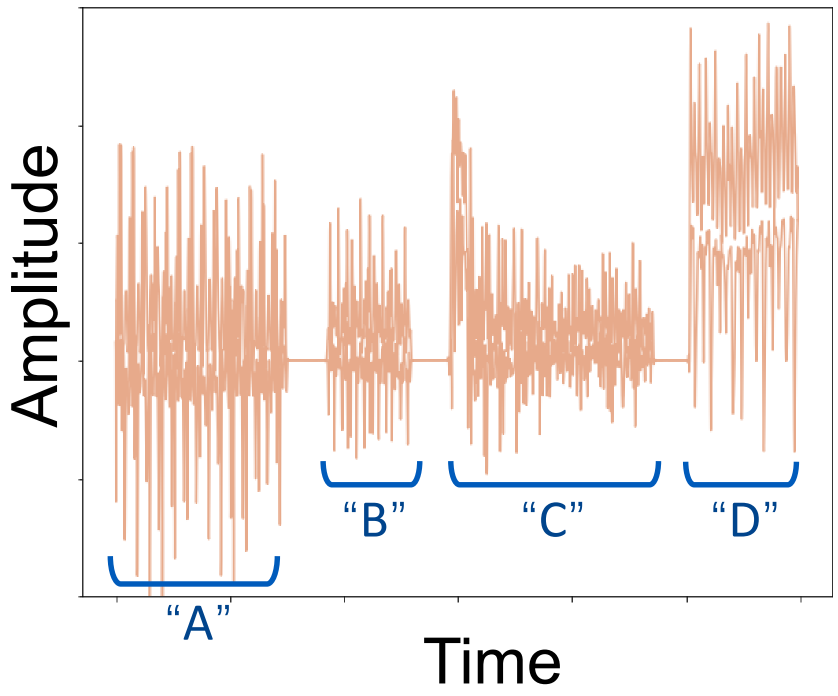


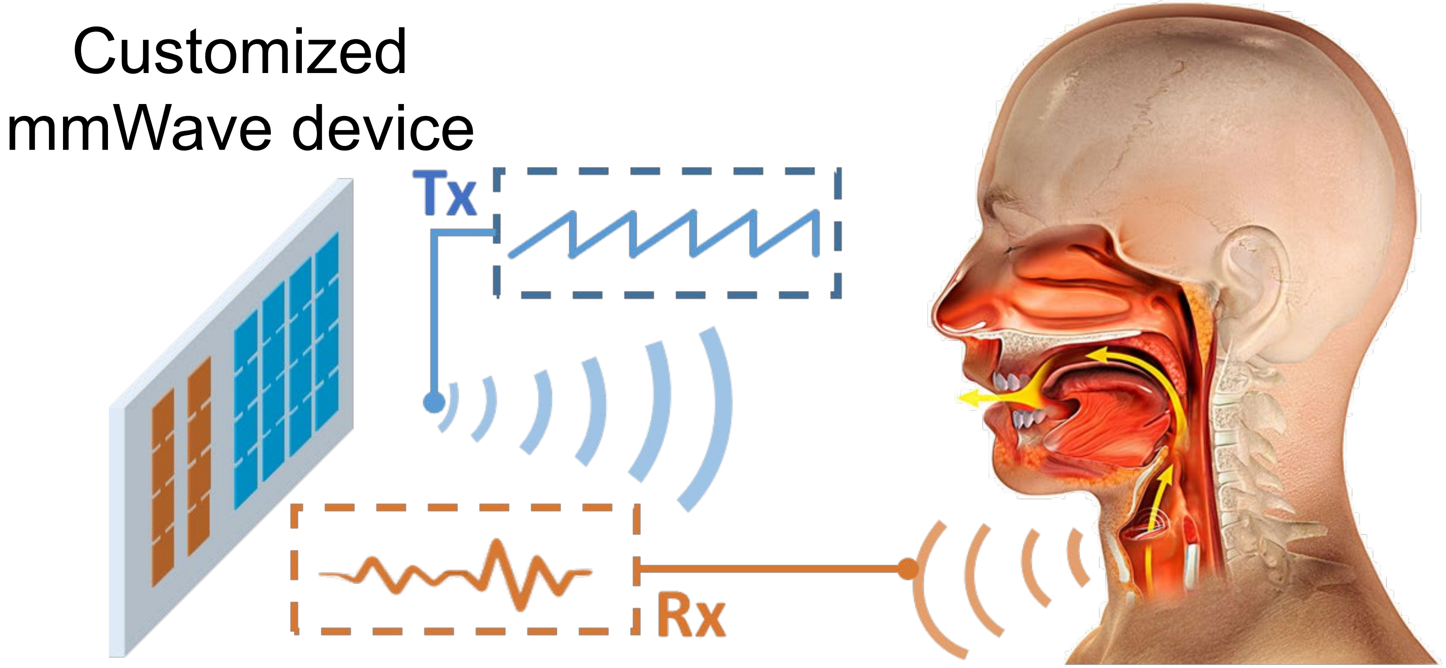
High-speed camera (SIGGRAPH'14)

Lidar sensors (SenSys'20)

# Near-throat vibration

- Recover speech information from human throat vibration

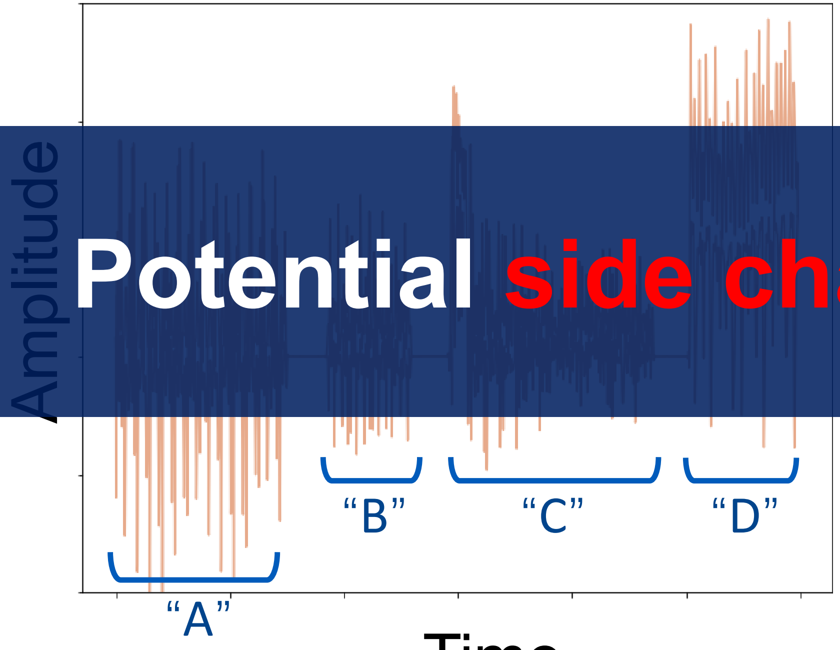



Xu, C., etc. (2019, June). Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services* (pp. 14-26).

# Near-throat vibration

- Recover speech information from human throat vibration



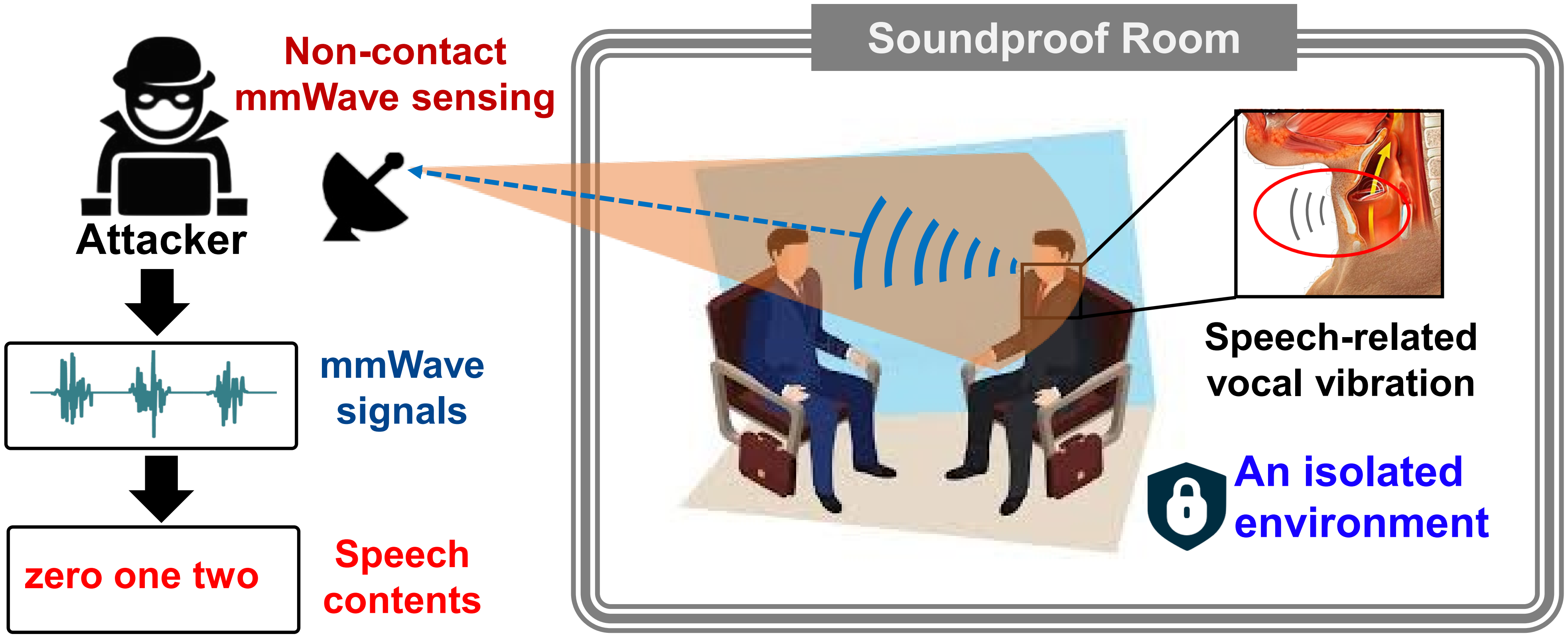

Customized mmWave device


**Potential side channel for confidential speech?**

Xu, C., etc. (2019, June). Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services* (pp. 14-26).

# Attack scenario

- Wireless-based through-wall eavesdropping



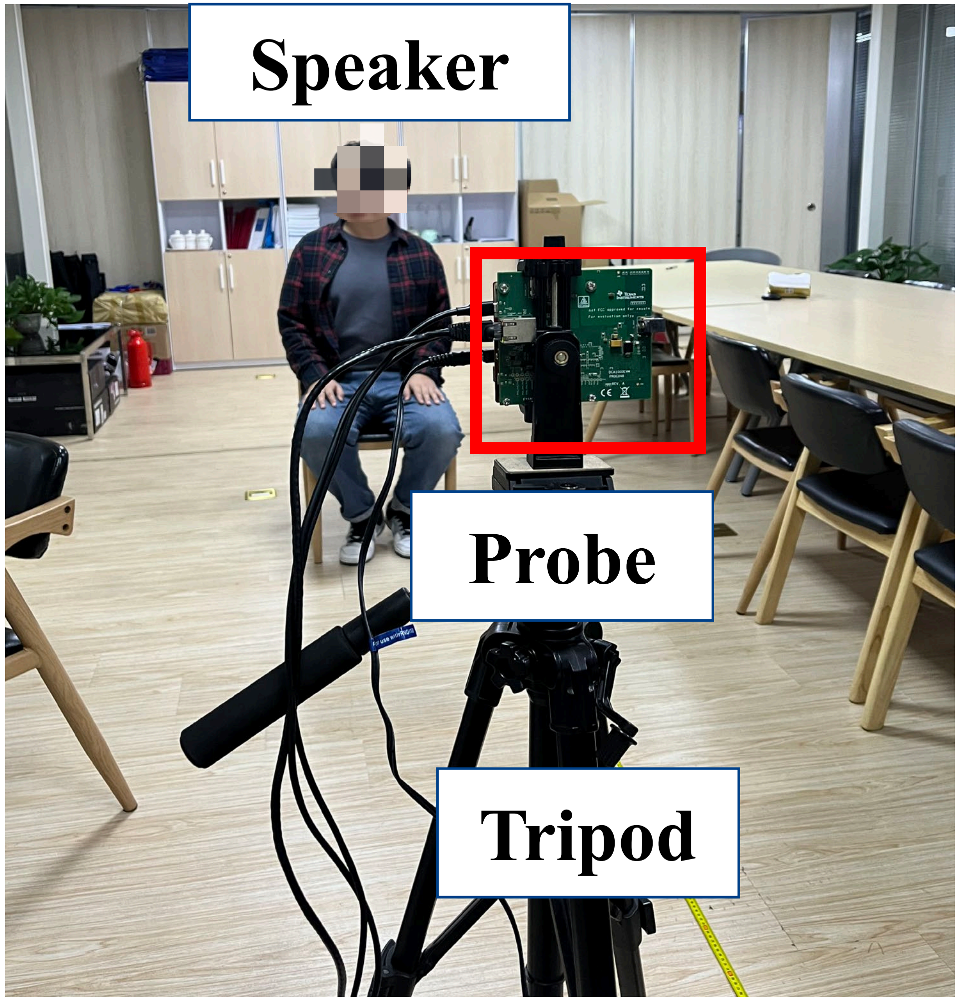
# Feasibility Study

**Experimental devices**

- mmWave probe (Tx/Rx)
- Laptop (data processing)

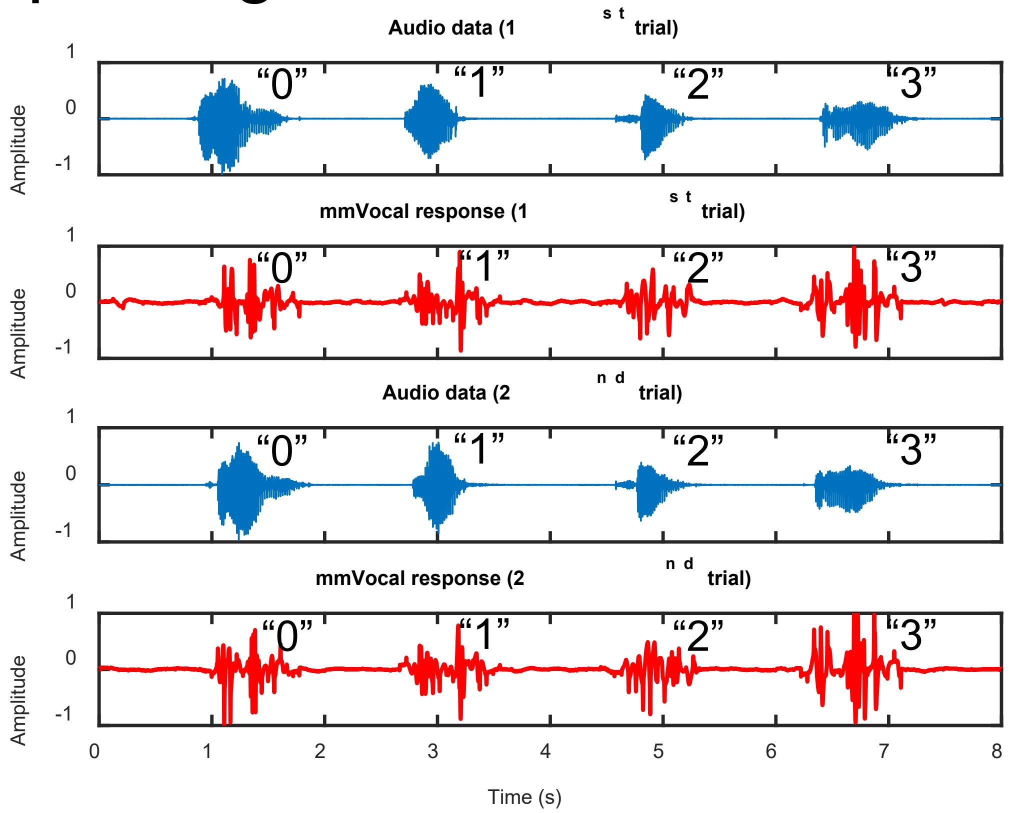
**Experimental setting**

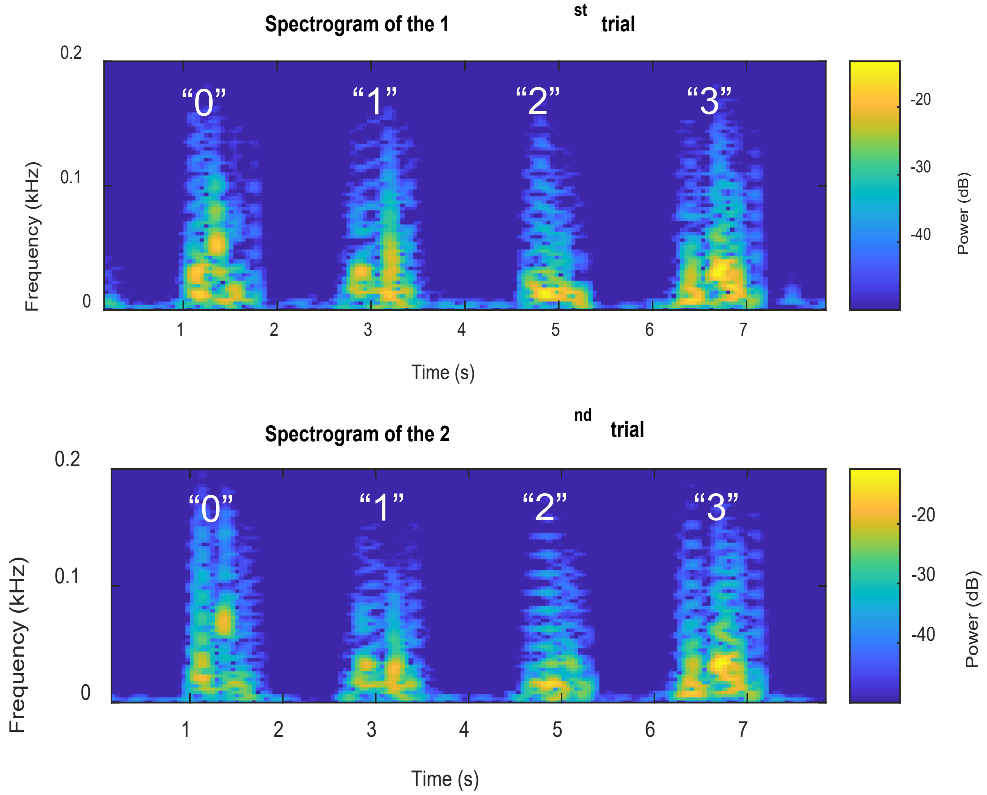
- Line-of-sight condition
- Distance: 2m
- “zero, one, two, three”



# Feasibility Study

- The patterns of the same speech shows a high similarity in the spectrogram.



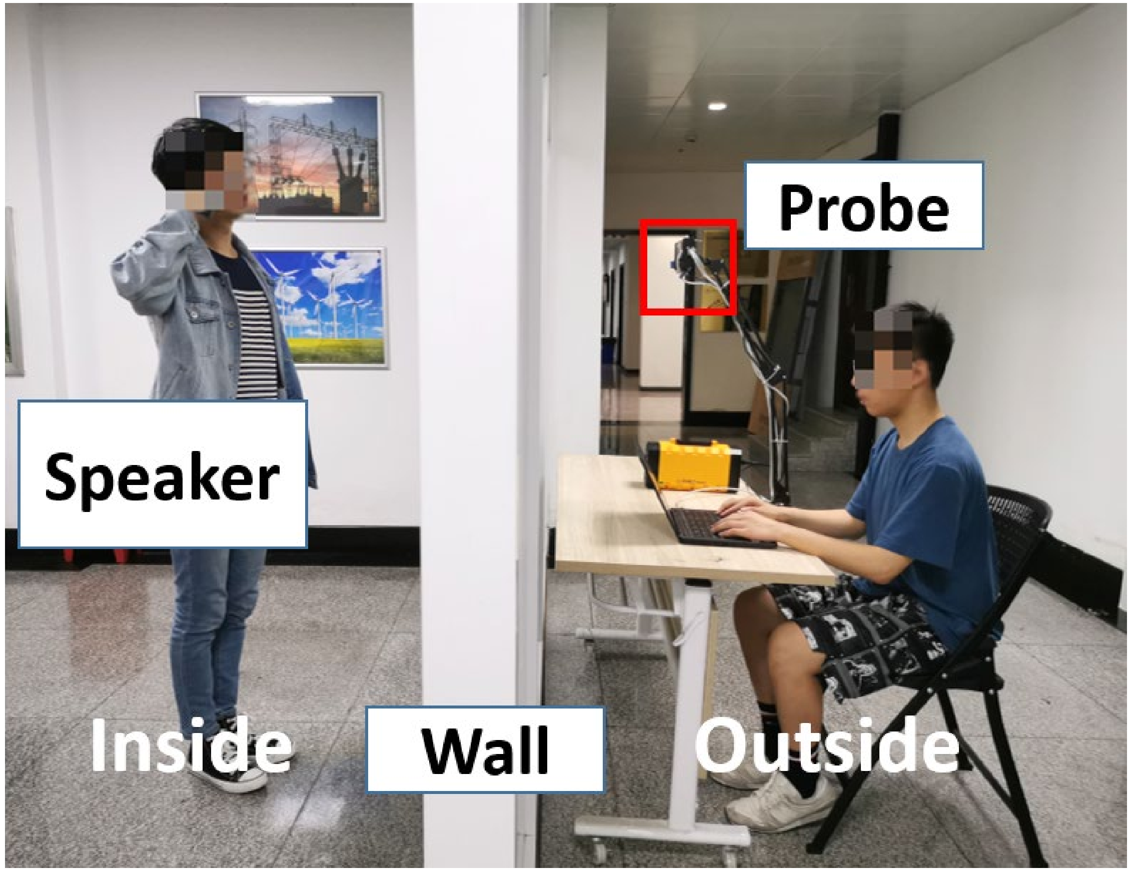
# NLoS condition

## Experimental devices

- mmWave probe (Tx/Rx)
- Laptop (data processing)

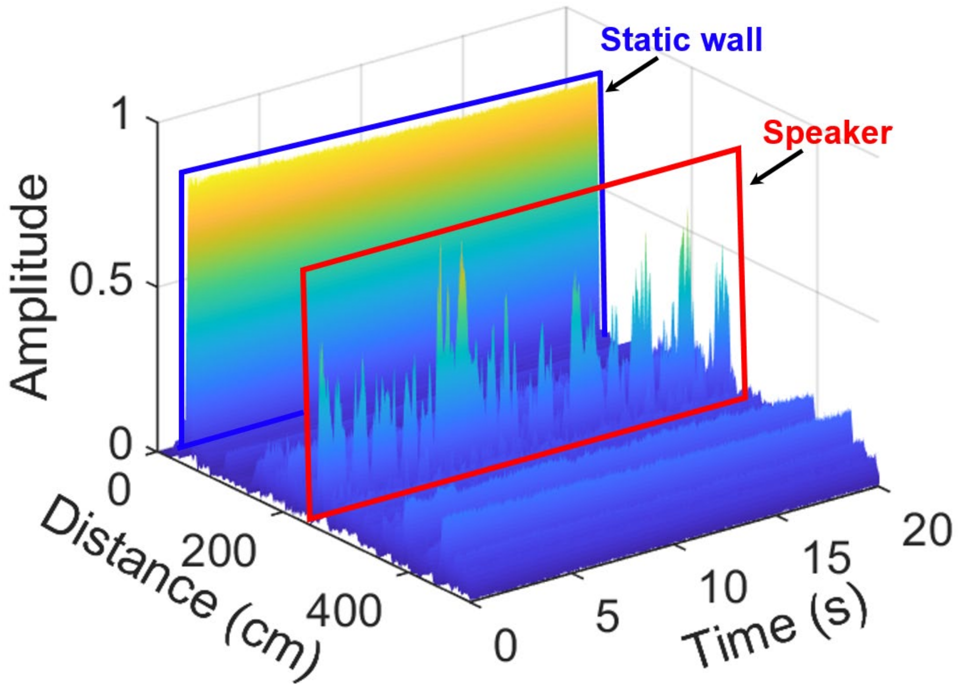
## Experimental Setting

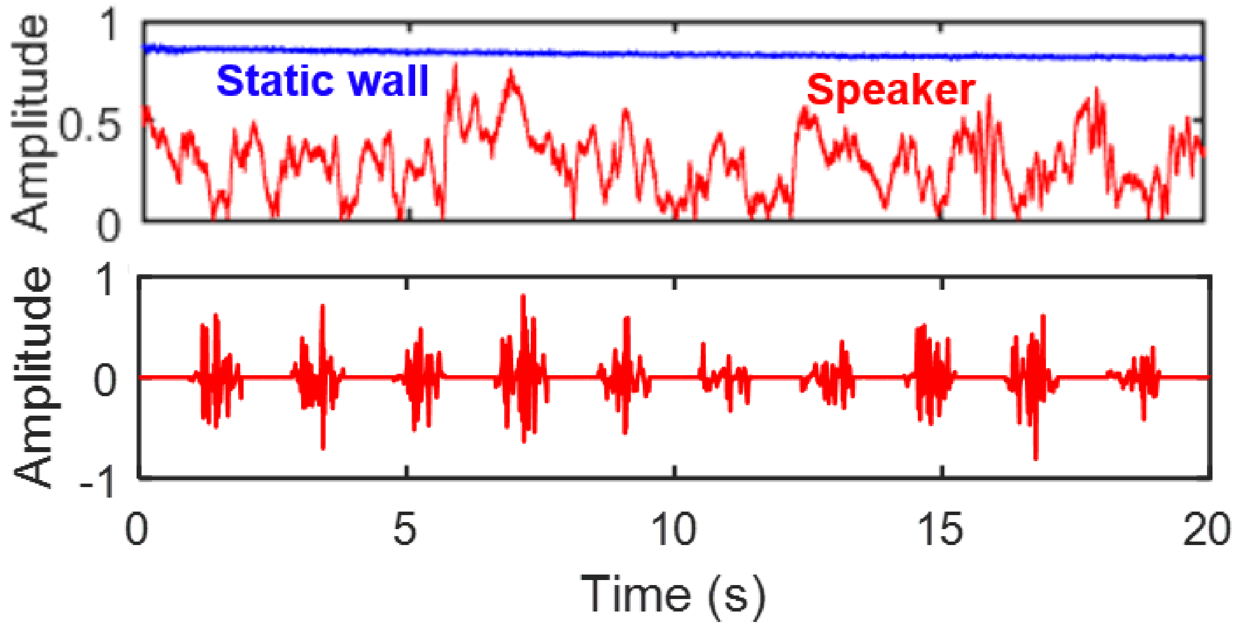
- Distance: 1m
- Through-wall sensing
- ”zero”…”nine”



# NLoS condition

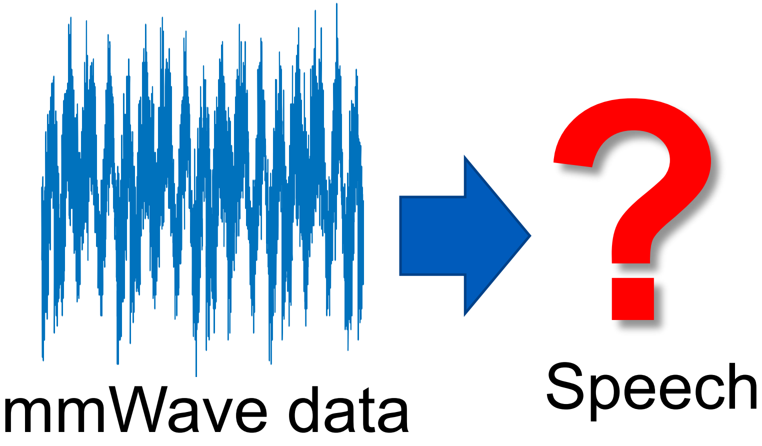
- The extracted vocal vibration is still observable in the time domain after wavelet analysis.



# Challenge






No prior knowledge about room layout

Background clutters

Implicit features for speech retrieval

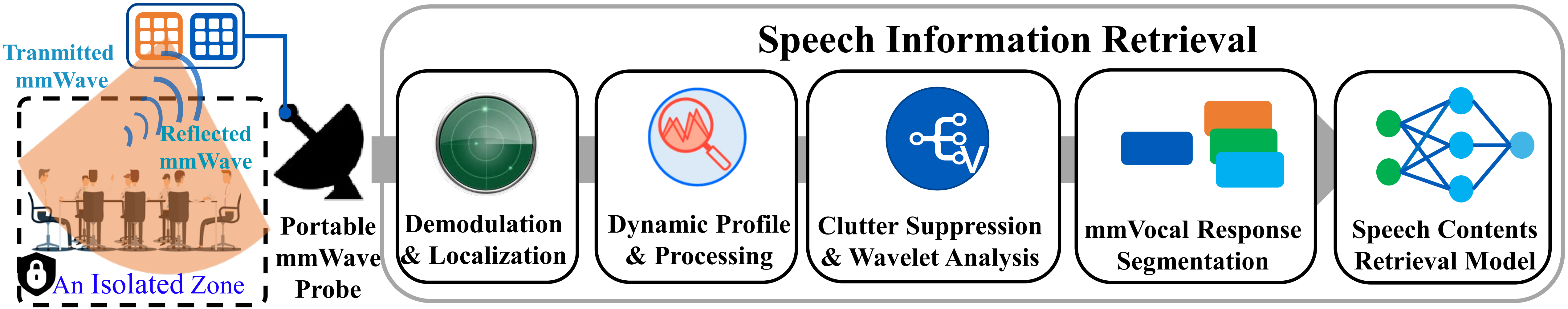
**Locate the speaker**

**Delicate vocal vibration**

**Speech retrieval**

# System Design

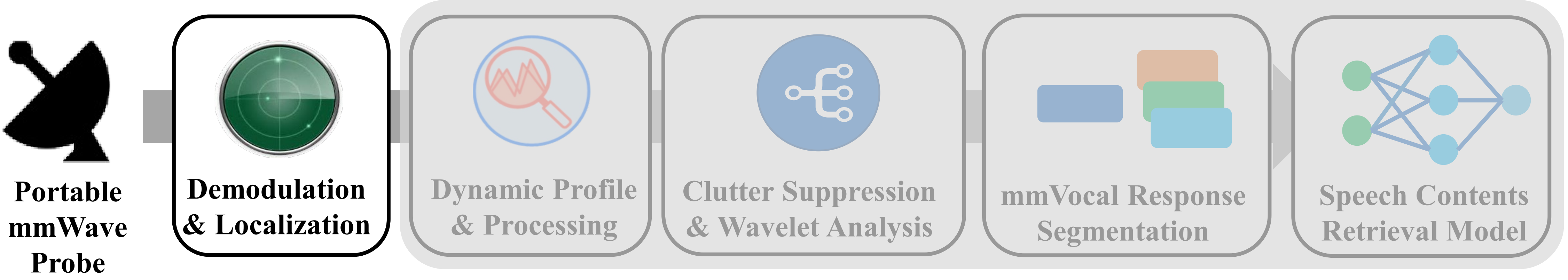
- **Wavesdropper: an end-to-end attack system**
  - **Through-wall** word detection via vocal vibration
  - Word recognition with **high accuracy**




Wavesdropper overview

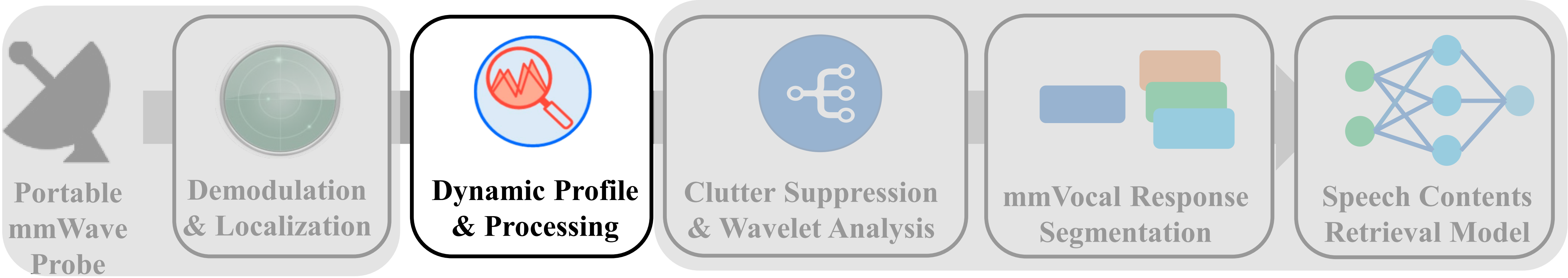
# Demodulation&Localization

- Range-FFT
  - Collect the demodulated data and apply FFT for each chirp
- Calculate the power density
  - Extract amplitude changes of every frequency point
  - The FFT point with the highest power density is selected



# Dynamic Profile&Processing

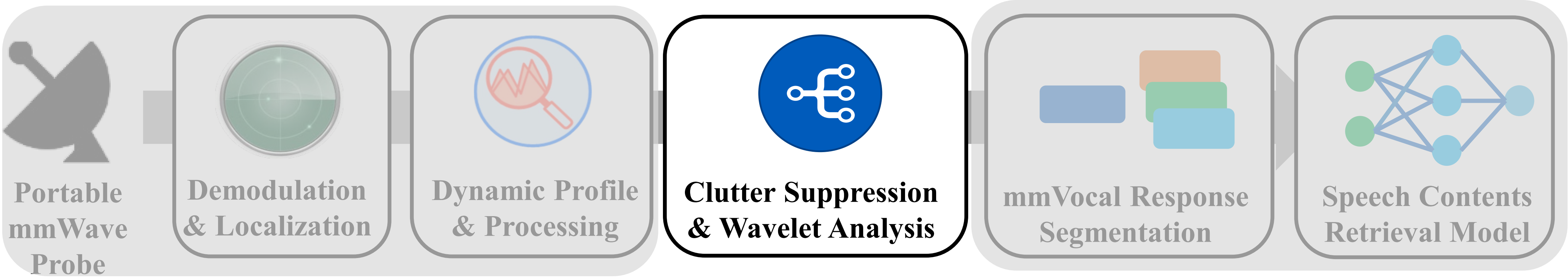
- Track the speaker
  - Apply the peak tracking on the spectrum after FFT
  - Extract the amplitude value of the selected frequency point
- Normalization
  - Constrain the amplitude within [-1,1]



# Clutter suppression&Wavelet analysis

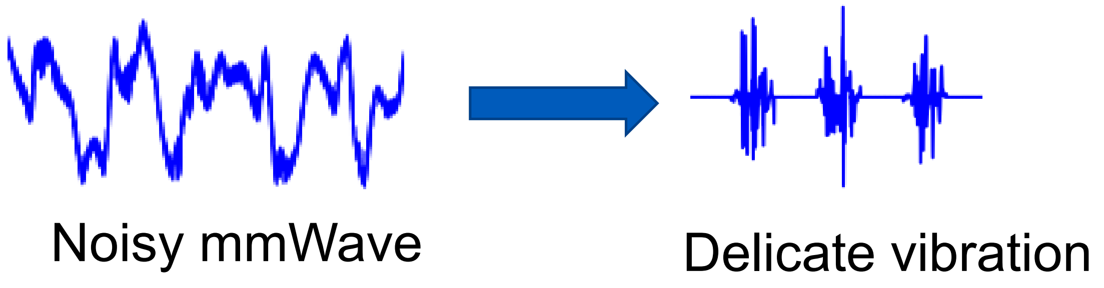
- CEEMD-based clutter suppression
  - CEEMD: $\{IMF_i(n)\} = CEEMD(s(n)), n = 1, \ldots N, i = 1, \ldots, I;$
  - Threshold-based reconstruction $T_r = \sigma_i \sqrt{2log(N)},$

$$IMF_i'(n) = \begin{cases} 0 & |IMF_i(n)| \le T_r \\ (2 * sigmoid(IMF_i) - 1)(|IMF_i| - T_r) & |IMF_i(n)| > T_r \end{cases}$$



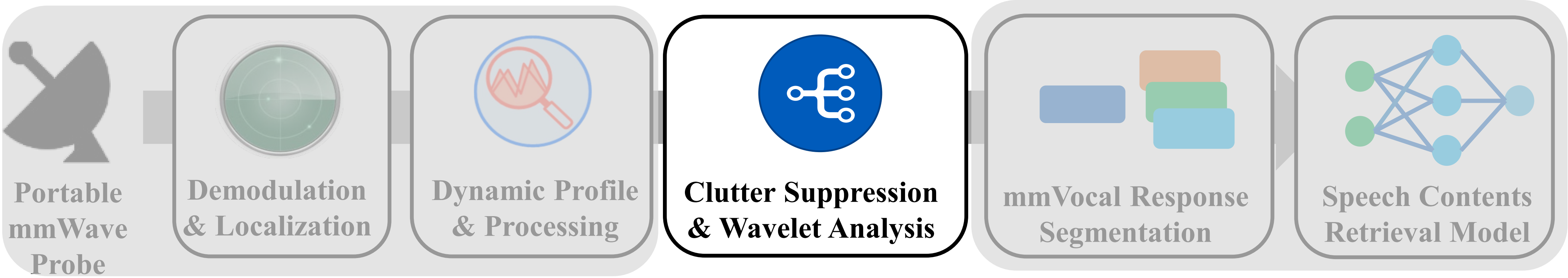
# Clutter suppression&Wavelet analysis

- Wavelet analysis
  - Wavelet Decomposition
  - Wavelet Reconstruction




$$s(t) = \boxed{A_0} + \boxed{D_1 + D_2 + D_3 + D_4 + D_5 + D_6}$$

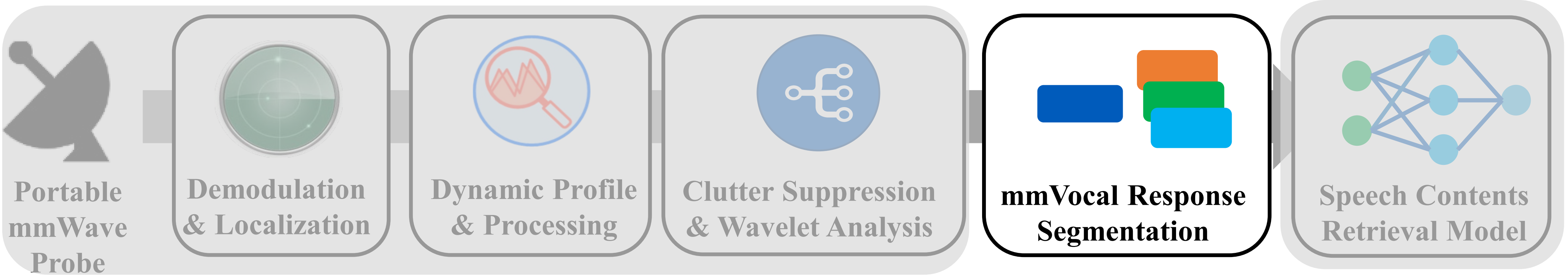
**Approximation part** **Detail part**



# mmVocal Response Segmentation

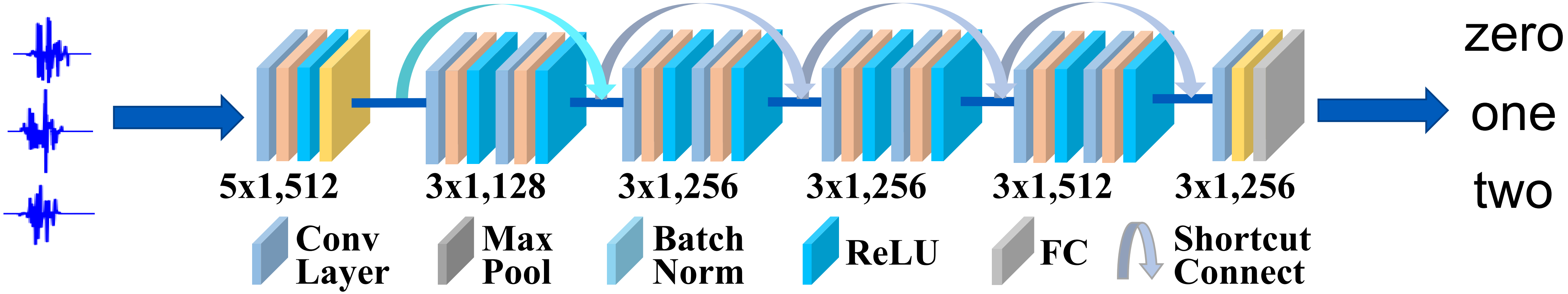
- Word-level segmentation
  - Segment the trace into several frames (50ms)
  - Calculate the signal energy *E* and spectral centroid *C* for each frame
  - Joint the successive frames with values higher than thresholds

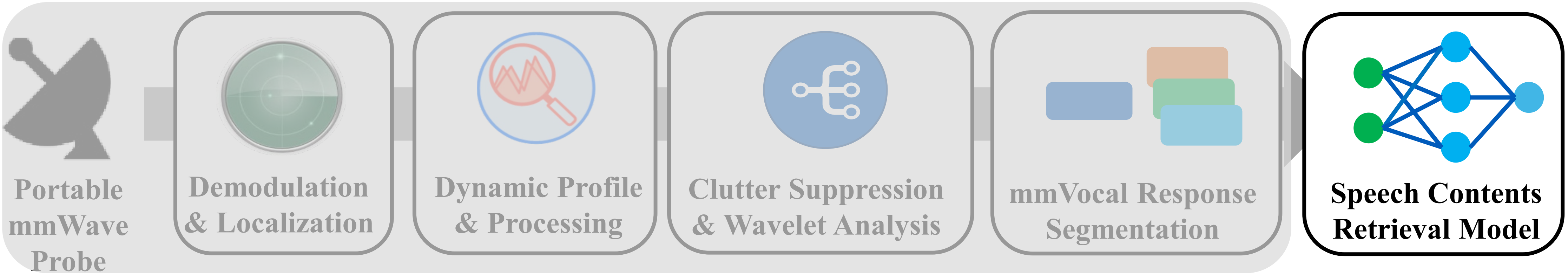
$$E(i) = \frac{1}{N}\sum_{n=1}^{N}|s_i(n)|^2 \qquad C(i) = \frac{\sum_{k=1}^{N}(k+1)S_i(k)}{\sum_{k=1}^{N}S_i(k)}$$



# Speech Retrieval

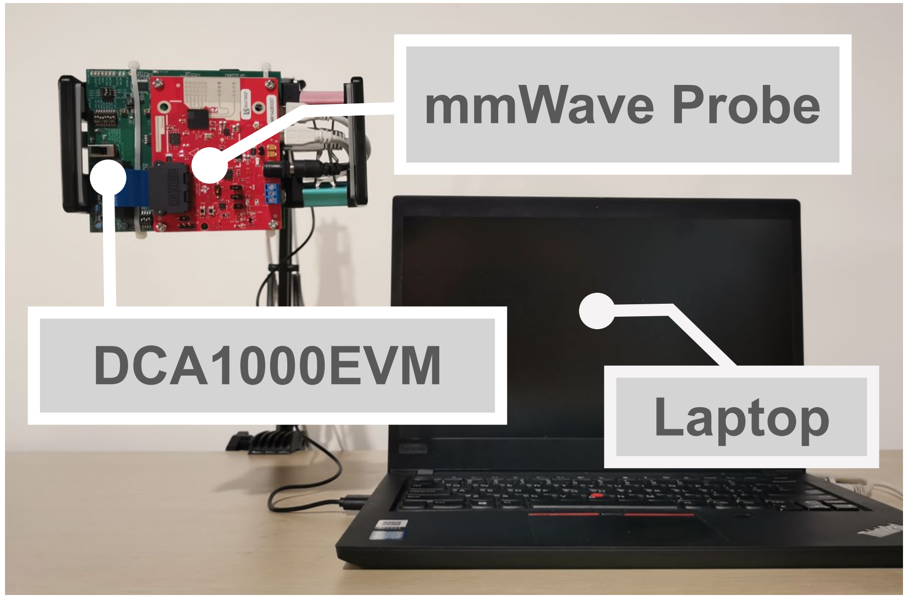
- DNN-based word recognition

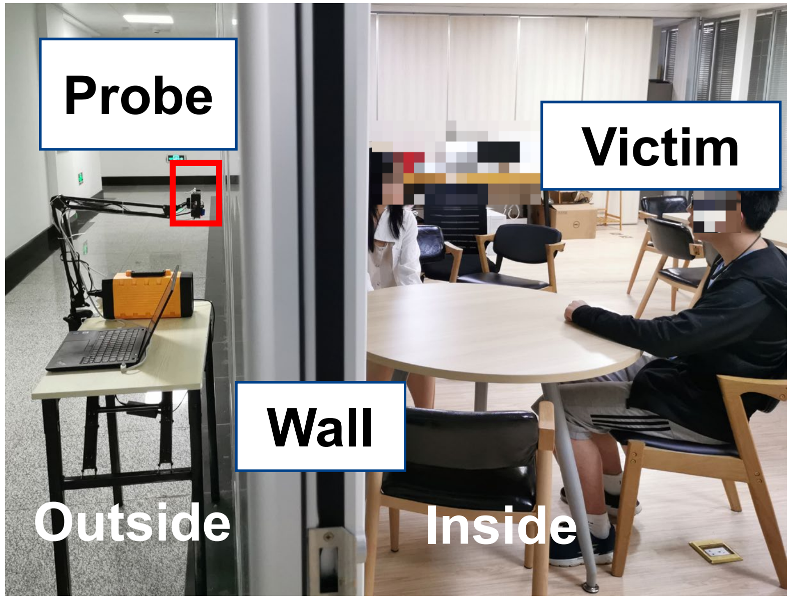


# Evaluation

- System setup
  - mmWave probe (IWR1642Boost) + DCA1000EVM
  - Laptop (Thinkpad 490) + Server (GeForce RTX 2060 GPU)
- Conference room with soundproof glasses



# Evaluation

- Dataset
  - 57 words (10 digits and 47 hot-words)
  - 23 volunteers (17 males and 6 females)
  - Over 50,000 samples in total
- Metrics
  - Top-k accuracy (Top-1, Top-3, Top-5)
  - mmVocal-Signal-to-Noise Ratio (mmVSNR)

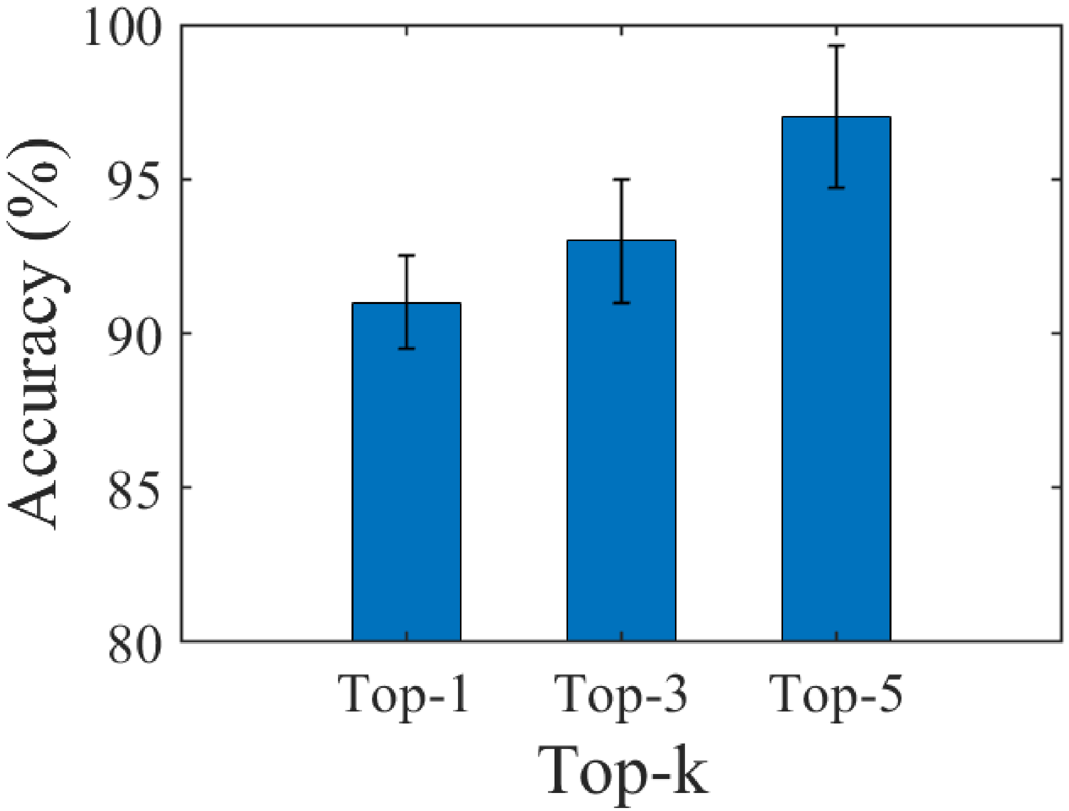
$$mmVSNR = 10\log_{10}\left(\frac{P(s)}{P(n)}\right)$$

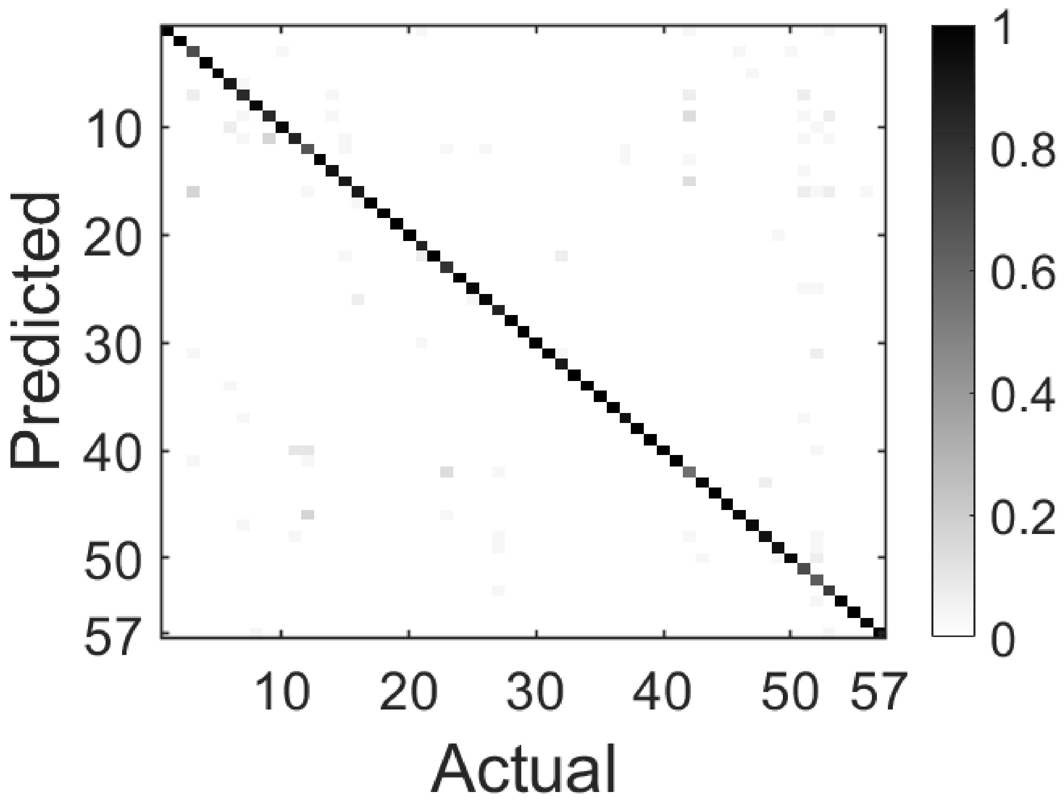
P(s) is the signal power of mmVocal response

P(n) is the signal power of the noise

# Overall performance

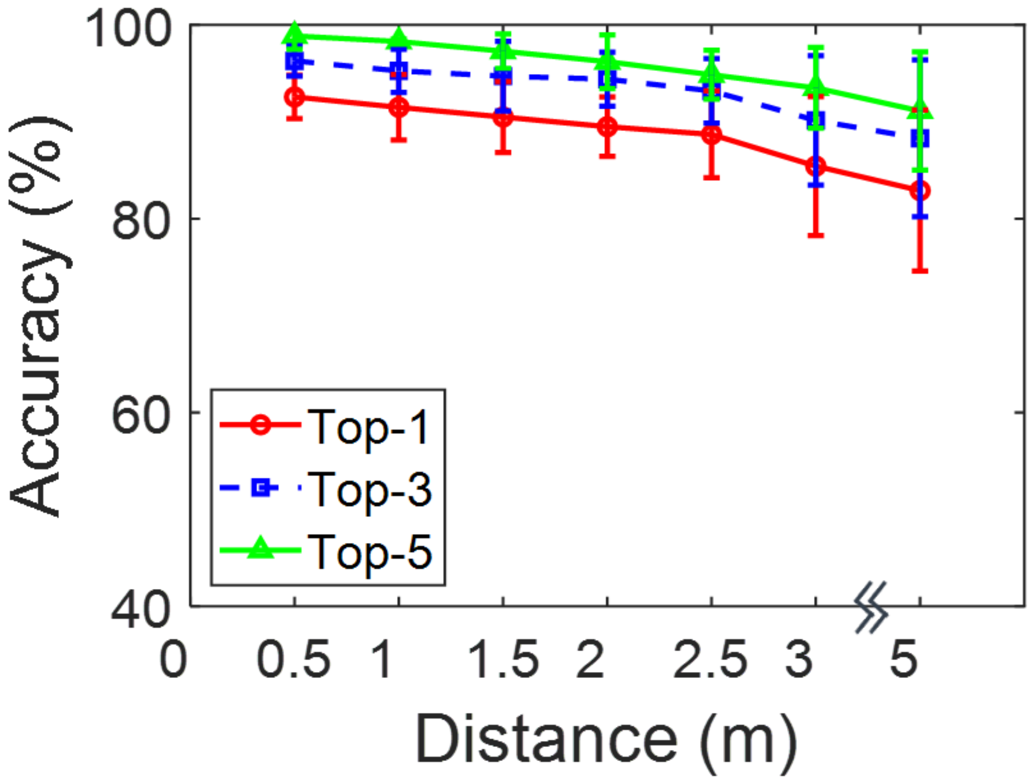
- Top-1 accuracy > 91% for 57-word recognition
- No obvious bias for the confusion matrix

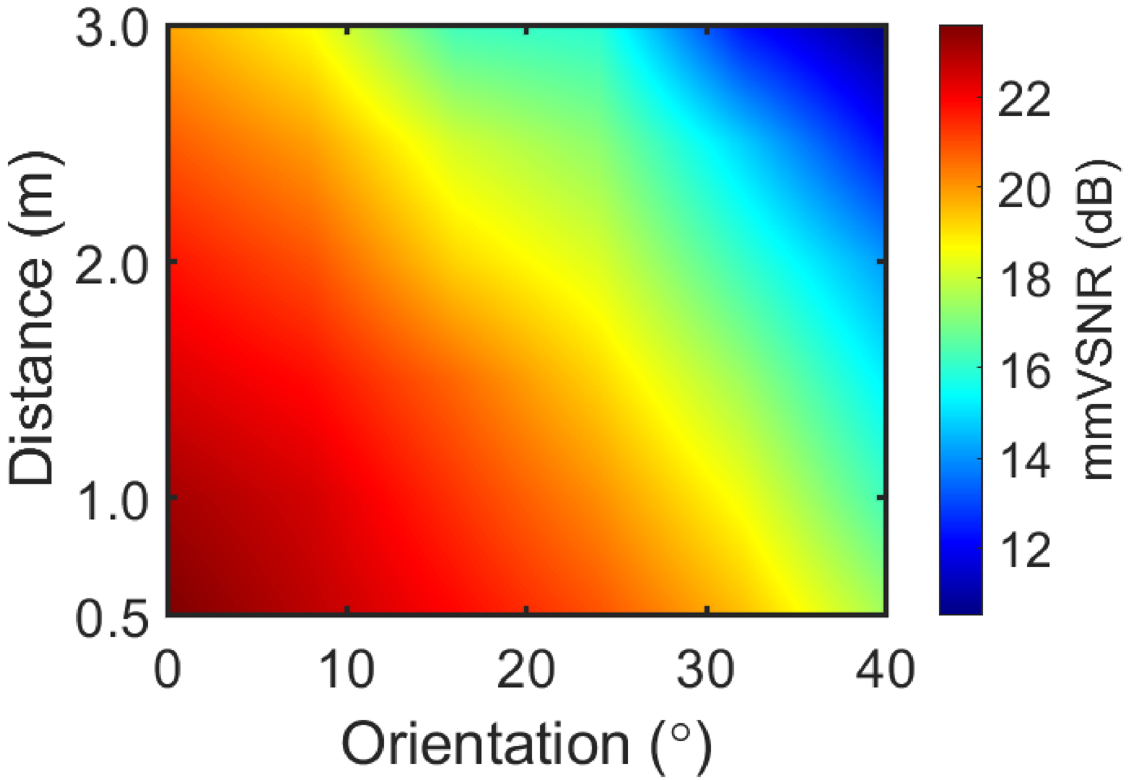


# Impact of distance&orientation

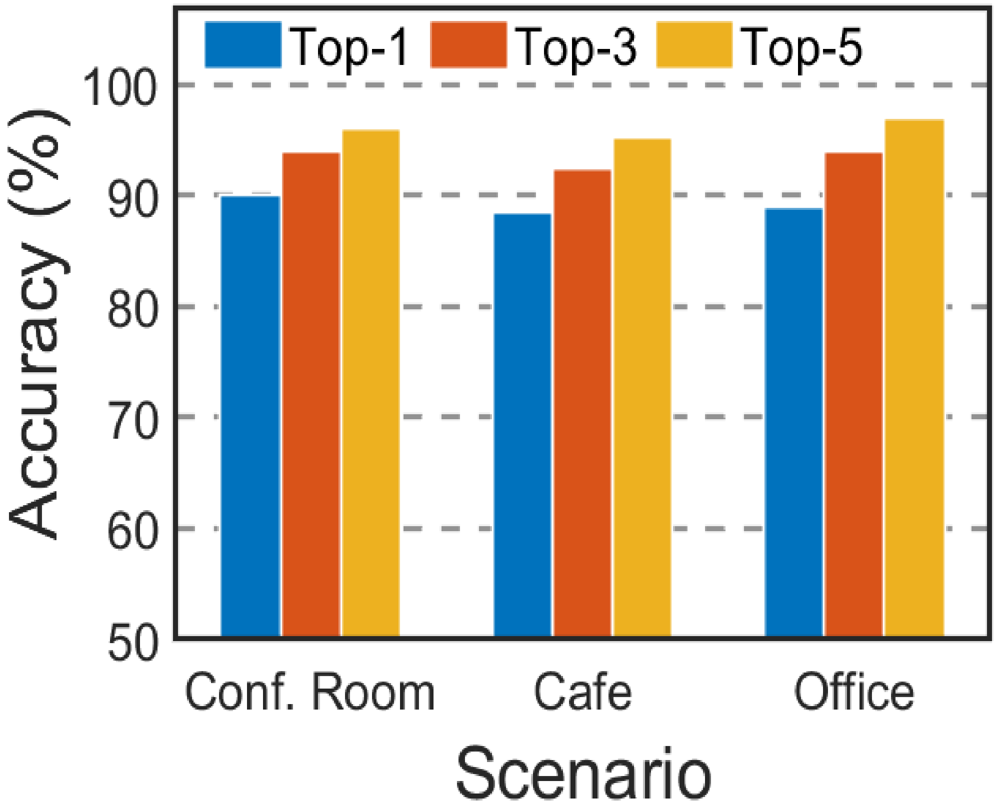
- Top-1 accuracy > 83% (distance < 5m)
- mmVSNR is stable (orientation < 30˚)

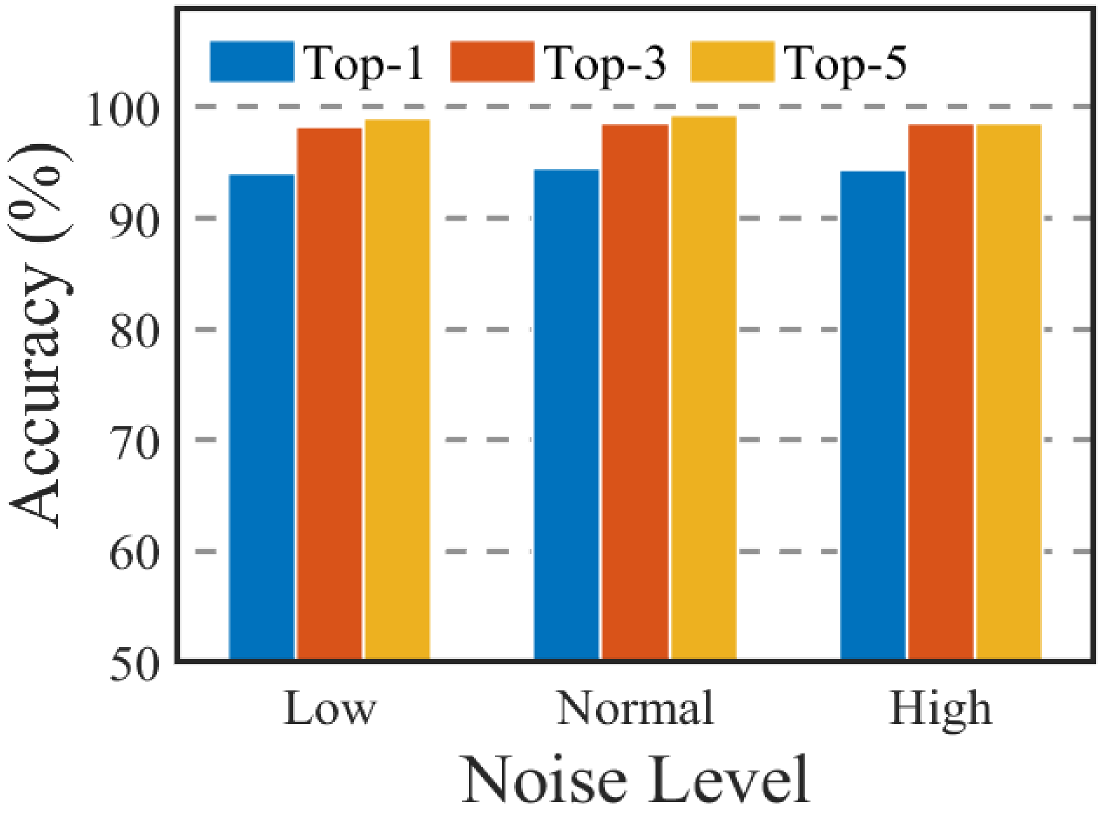


# Environment changes &Background noise

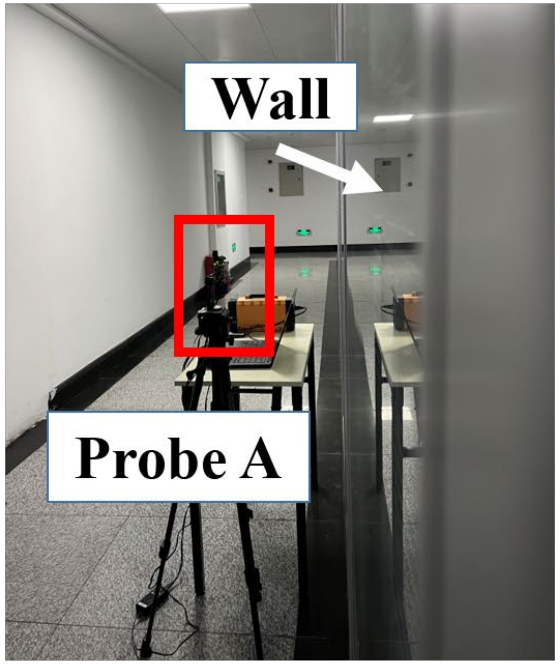
- Resilient to environment changes and background noise



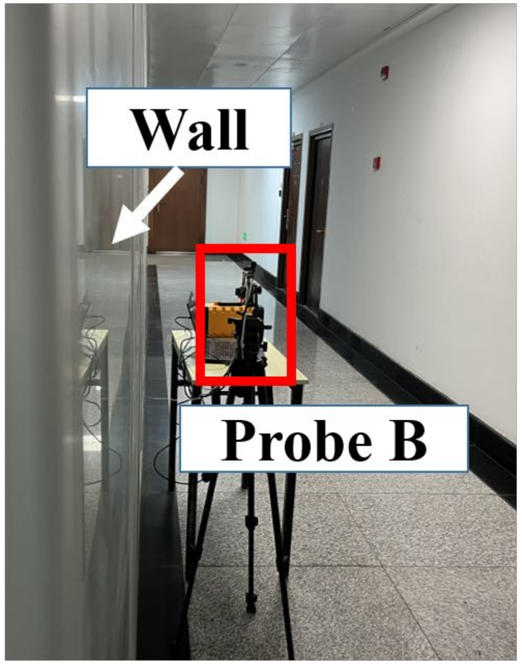
# Word detection on two speakers

- Two probes for through-wall word detection
- Both of the two volunteers: Top-1 accuracy > 88%




**Outside**

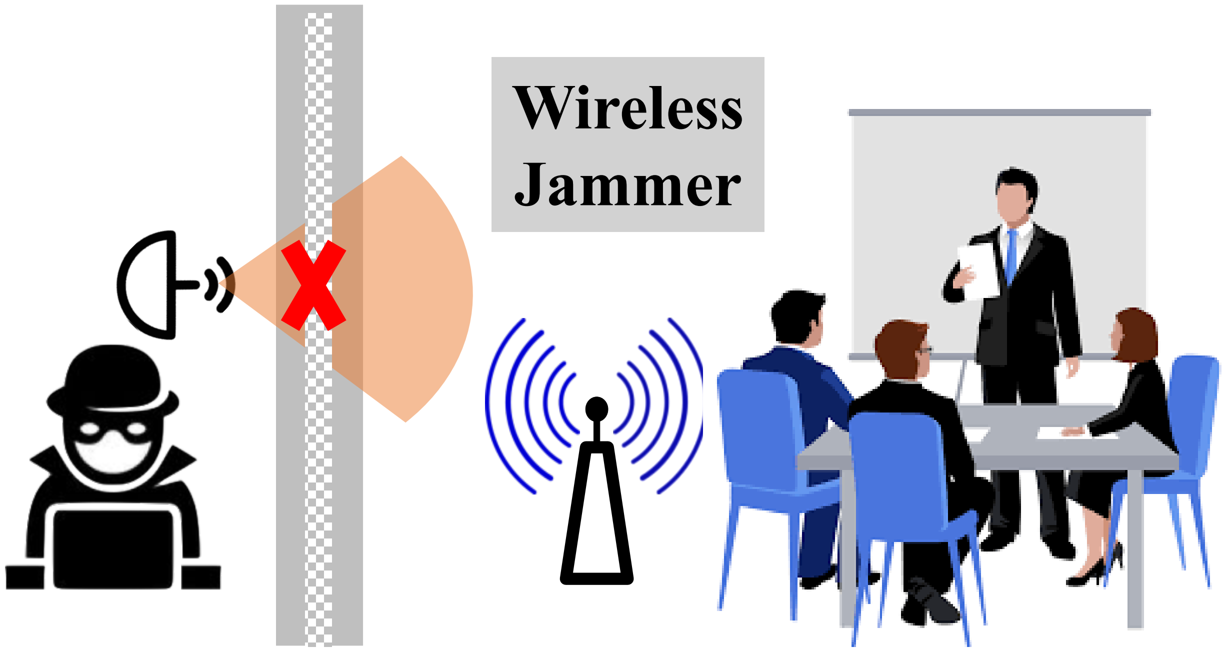

**In-room**

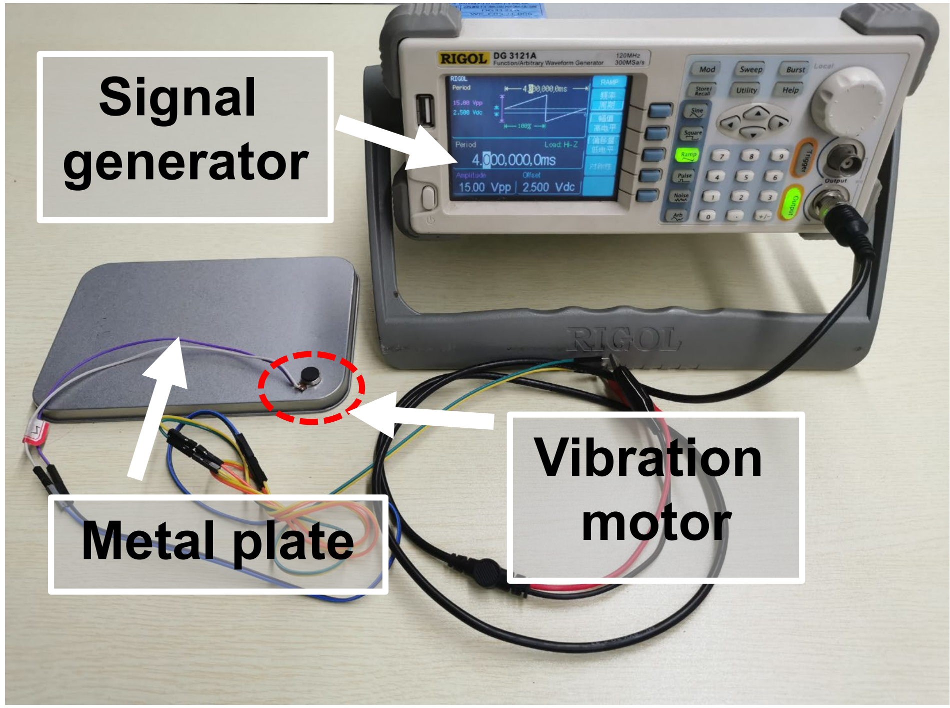



**Outside**

# Countermeasure

- Shielding or Jamming the malicious mmWave signals
- Interfering vocal vibration with another vibration source



# Conclusion

- A new speech threat
  - Through-wall word detection
  - Commercial off-the-shelf mmWave devices
- An end-to-end attack system
  - Recognize up to 57 words with high accuracy
- Countermeasures
  - Shielding&Jamming
  - Vibration interference

# Thanks for listening!