The Catcher in the Field: A Fieldprint based Spoofing Detection for Text-Independent Speaker Verification

Chen Yan, Yan Long, Xiaoyu Ji, Wenyuan Xu

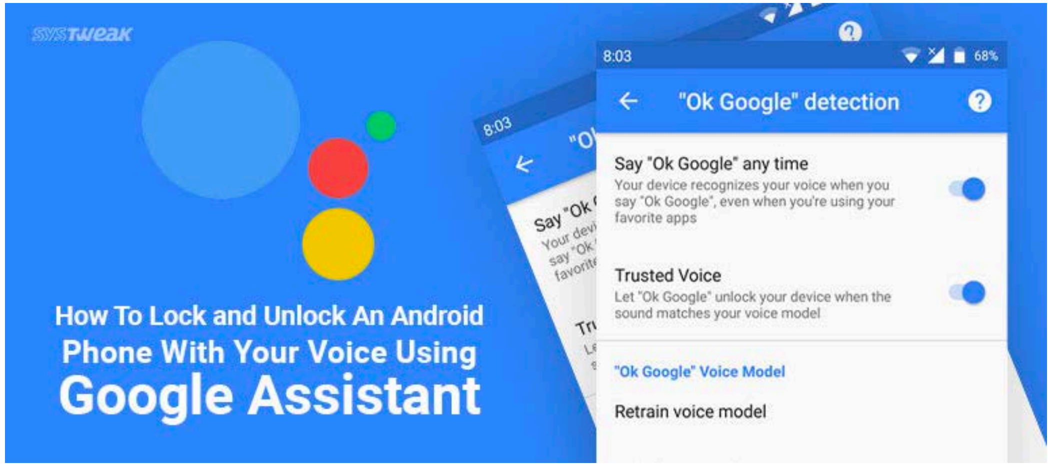
Zhejiang University

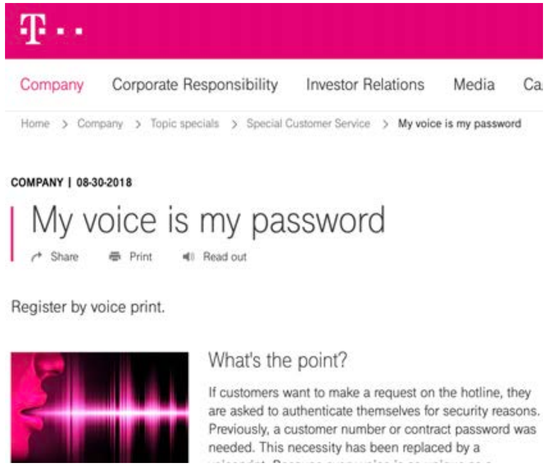
浙江大學
ZHEJIANG UNIVERSITY

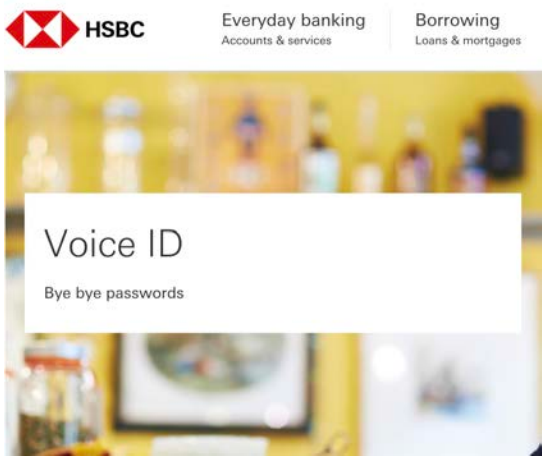
智能系统安全实验室
UBIQUITOUS SYSTEM SECURITY LAB.

Voice biometrics - “My voice is my identity”

- ❑ Unique human voice -> Identity
- ❑ Speaker recognition & verification
- ❑ Device unlock, voice assistant, account login, banking, …







Is voice biometrics as **sound** as it sounds?

Voice can be faked by attackers

THE WALL STREET JOURNAL.

Subscrib

English Edition ▾ | October 30, 2019 | Print Edition | Video

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

BREAKING NEWS Pace of U.S. economic growth slowed slightly to 1.9% in third quarter as business investment declined, though consumer spending kept growth on track

PRO CYBER NEWS

SHARE

AA TEXT

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

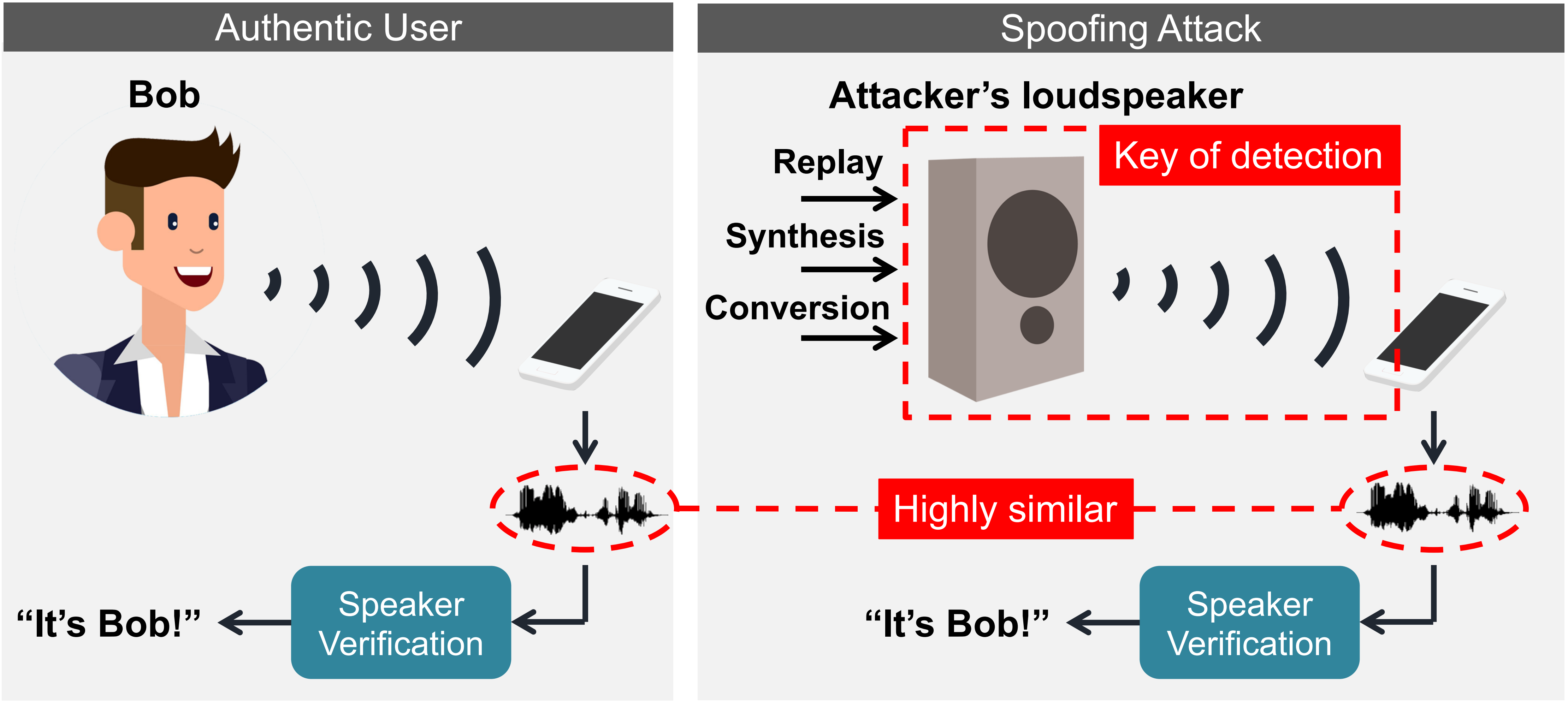
- ❑ Voice spoofing attacks:
  - Replay
  - Voice synthesis
  - Voice conversion
- ❑ Attacks made easier with off-the-shelf tools

Lyrebird DeepMind Adobe

Our goals

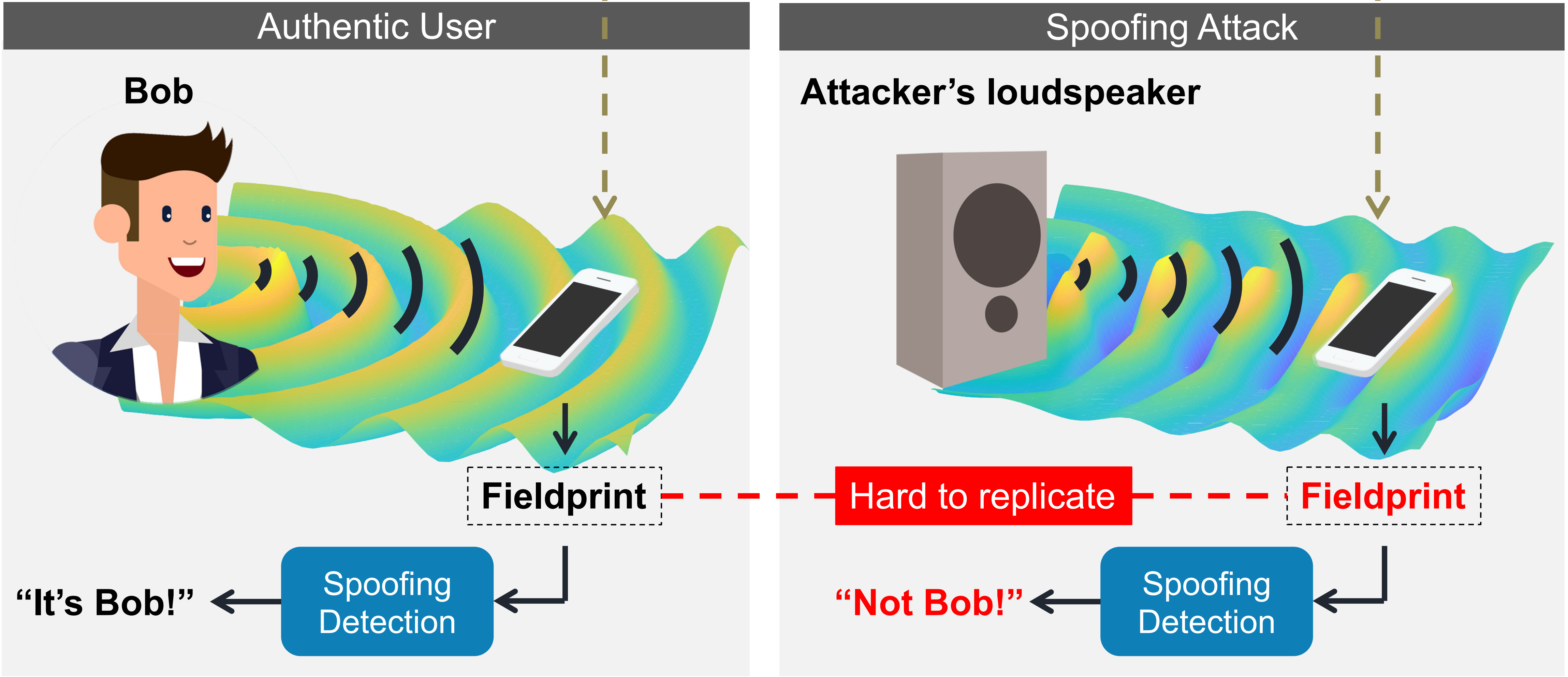
- ❑ Detect voice spoofing attacks
- ❑ Applicable to smartphones
- ❑ Balance **security** & **usability**
  - Text-independent
  - No extra device
  - User-defined device positions

Key insight 1



Key insight 2

A **sound field** describes the dispersion of acoustic energy over space

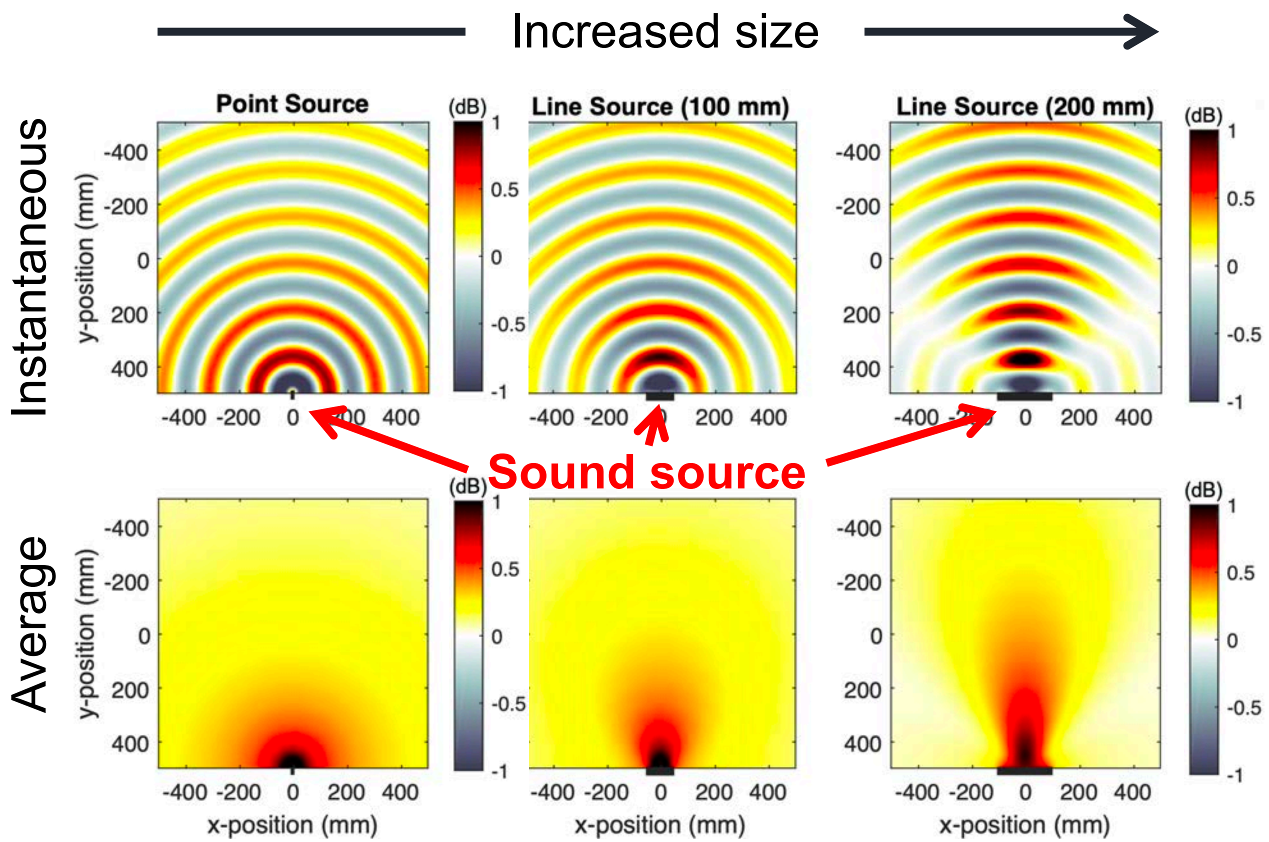


Research questions

Q1 **Can the sound fields of authentic users and spoofing attackers be different?**

Q1 Simulation of sound fields in MATLAB

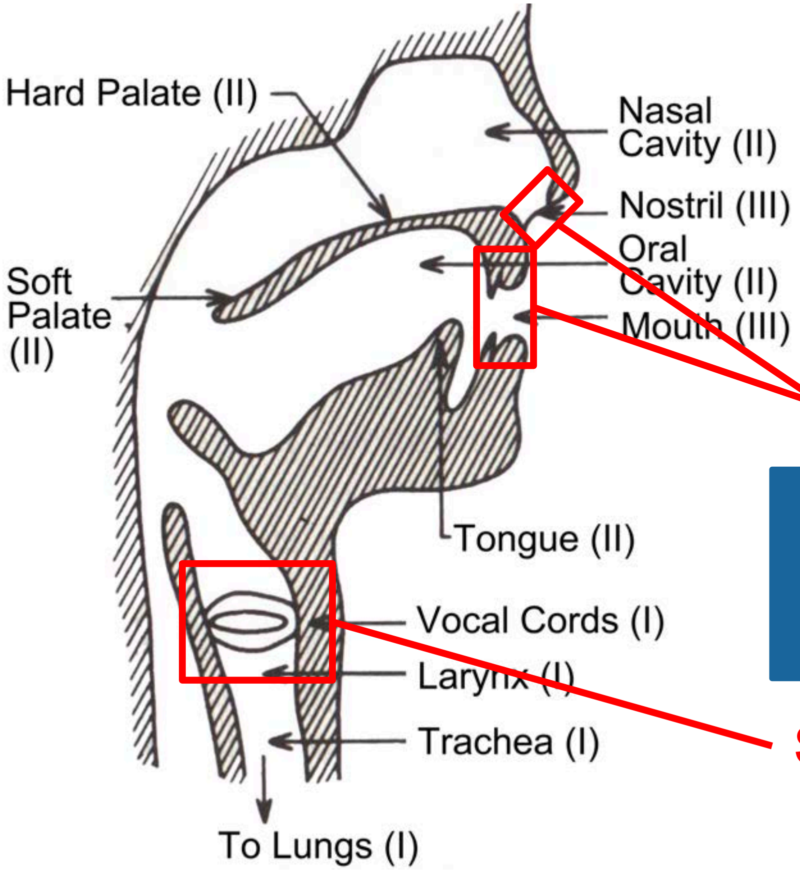
(A) Effect of the size of the sound source



Larger size → More directional

What is the difference of human and loudspeaker in size as sound sources?

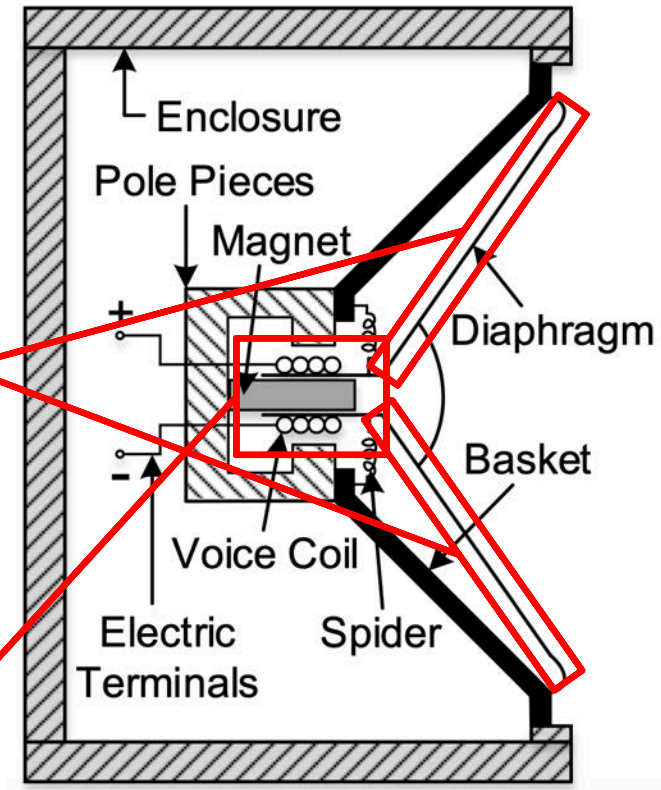
Q1 Human and loudspeaker in size



Parts of radiation

Creates sound field;
Essentially different
in size and shape

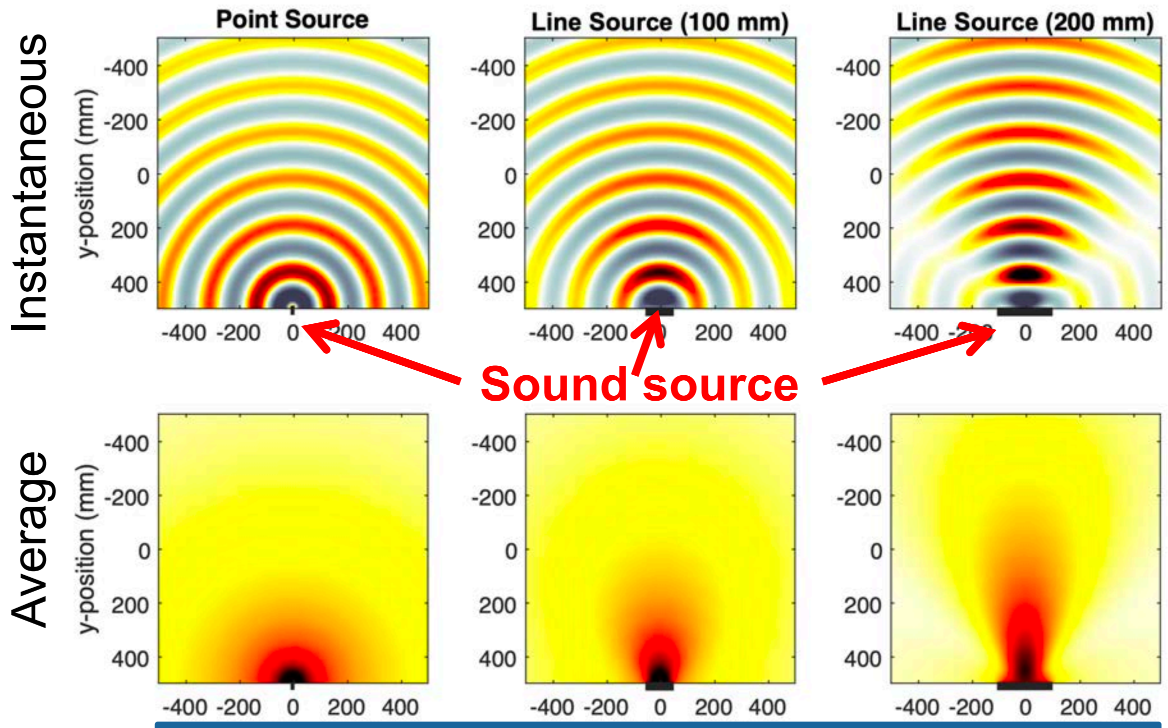
Source of vibration



Q1 Simulation of sound fields in MATLAB

(A) Effect of the size of the sound source

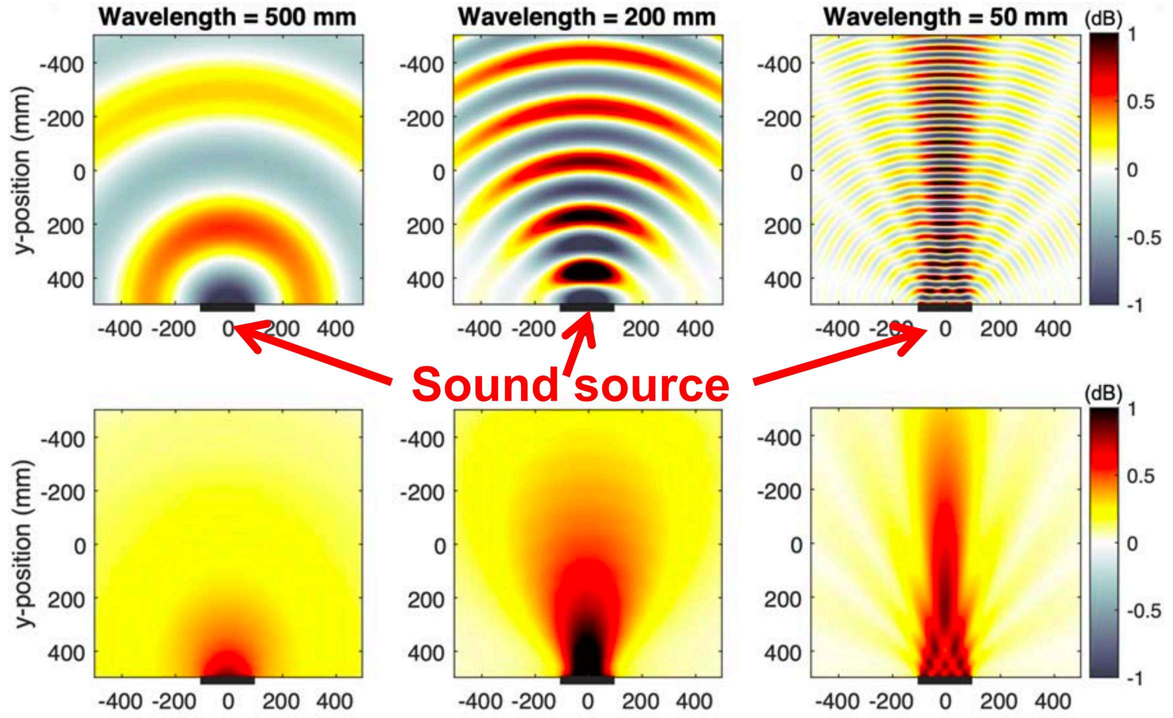
Increased size



Larger size → More directional

(B) Effect of the sound wavelength

Decreased wavelength



Higher freq. → More directional

Research questions

Q1 Can the sound fields of authentic users and spoofing attackers be different?

Q2 **How to extract fieldprints from sound fields without using devices other than a smartphone?**

Q2 Fieldprint formulation

- Limited number of microphones on a smartphone (mostly 2~3)
- Difference of acoustic energy (sound frequency f) at the 2 microphone locations ($\boldsymbol{p_1}, \boldsymbol{p_2}$):

$$S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f) = \log \frac{S(\boldsymbol{p_1}, f)}{S(\boldsymbol{p_2}, f)}$$

← Sound pressure at Mic 1
← Sound pressure at Mic 2

- Basic fieldprint:

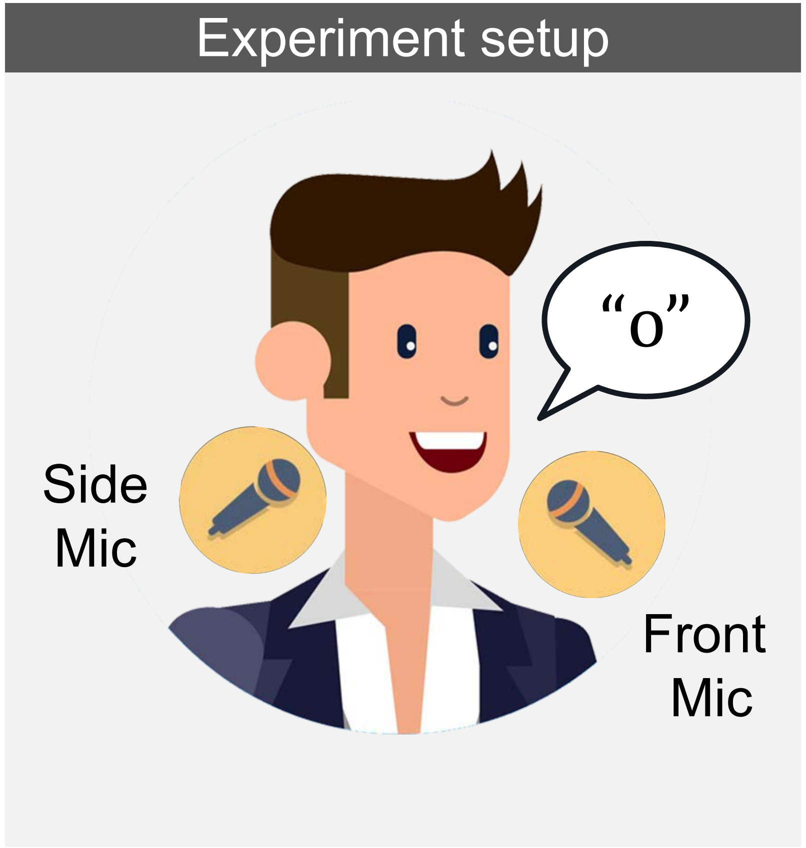
$$\mathcal{F}(\boldsymbol{p_1}, \boldsymbol{p_2}) = [S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f_1), S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f_2), \ldots, S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f_n)]$$

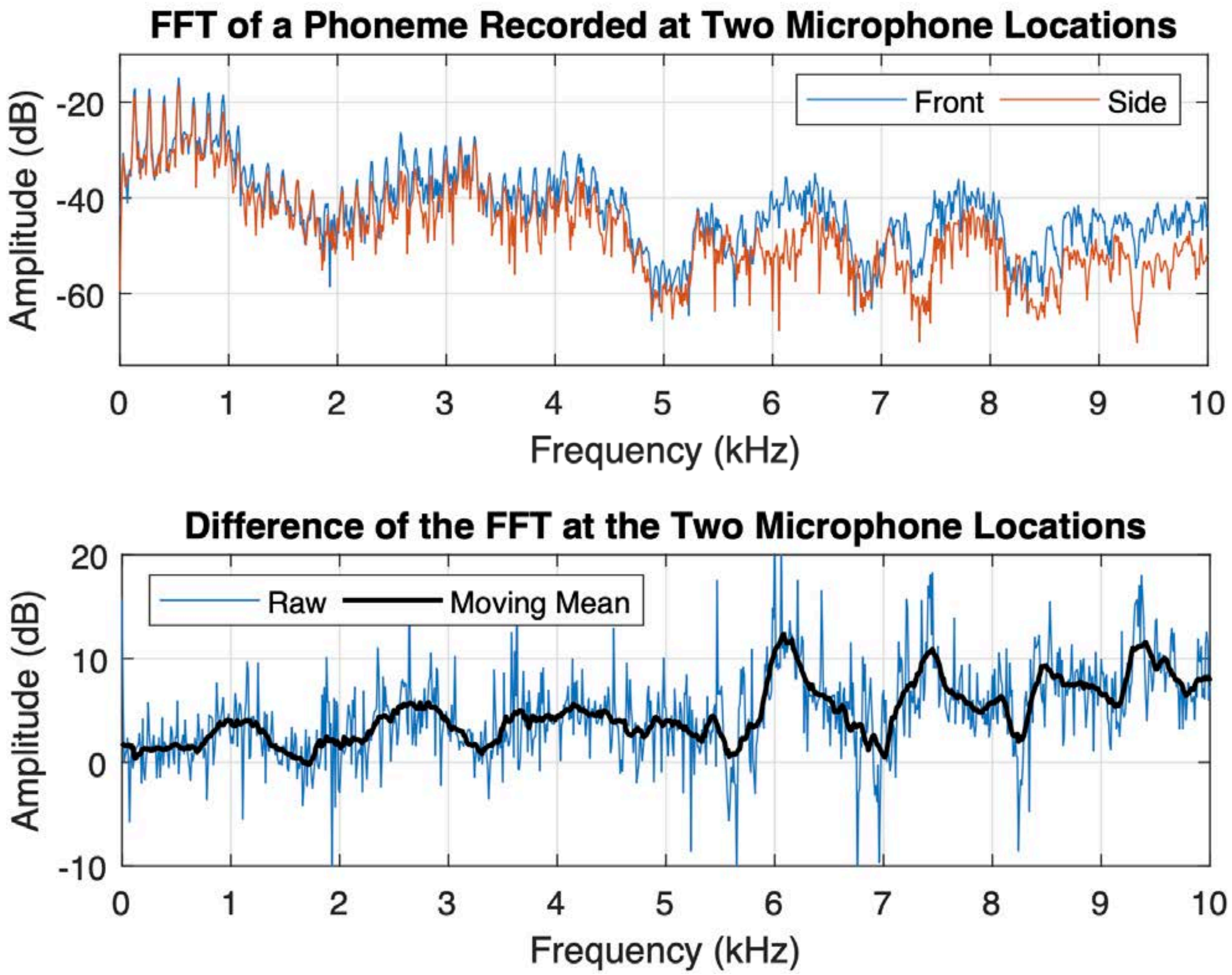
Q2 Fieldprint formulation

- Basic fieldprint:

$$
\begin{aligned}
\mathcal{F}(\boldsymbol{p_1},\boldsymbol{p_2}) &= [S_R(\boldsymbol{p_1},\boldsymbol{p_2},f_1), S_R(\boldsymbol{p_1},\boldsymbol{p_2},f_2), \dots S_R(\boldsymbol{p_1},\boldsymbol{p_2},f_n)] \\
&= \left[\log\frac{S(\boldsymbol{p_1},f_1)}{S(\boldsymbol{p_2},f_1)}, \log\frac{S(\boldsymbol{p_1},f_2)}{S(\boldsymbol{p_2},f_2)}, \dots, \log\frac{S(\boldsymbol{p_1},f_n)}{S(\boldsymbol{p_2},f_n)}\right] \\
&= [\log\big(S(\boldsymbol{p_1},f_1)\big) - \log\big(S(\boldsymbol{p_2},f_1)\big), \log\big(S(\boldsymbol{p_1},f_2)\big) - \log\big(S(\boldsymbol{p_2},f_2)\big), \\
&\quad \dots, \log\big(S(\boldsymbol{p_1},f_n)\big) - \log\big(S(\boldsymbol{p_2},f_n)\big)] \\
&= [\log\big(S(\boldsymbol{p_1},f_1)\big),\ \log\big(S(\boldsymbol{p_1},f_2)\big), \dots, \log\big(S(\boldsymbol{p_1},f_n)\big)] \\
&\quad -[\log\big(S(\boldsymbol{p_2},f_1)\big), \log\big(S(\boldsymbol{p_2},f_2)\big), \dots, \log\big(S(\boldsymbol{p_2},f_n)\big)] \\
&\boxed{= \log\big(\mathrm{FFT}(<\text{sound at } \boldsymbol{p_1}>)\big) - \log\big(\mathrm{FFT}(<\text{sound at } \boldsymbol{p_2}>)\big)}
\end{aligned}
$$

Q2 Fieldprint formulation - Benchmark experiment





Research questions

Q1 Can the sound fields of authentic users and spoofing attackers be different?

Q2 How to extract fieldprints from sound fields without using devices other than a smartphone?

Q3 **To what degree do fieldprints show consistency and distinctiveness?**

Q3 Fieldprint consistency and distinctiveness

❑ Consistency

- The ability to be consistent
- Text-independent: effect of the speech content
- Microphone location: effect of the microphone locations

❑ Distinctiveness

- The ability to be distinctive between human and loudspeakers
- The ability to be distinctive between different people?

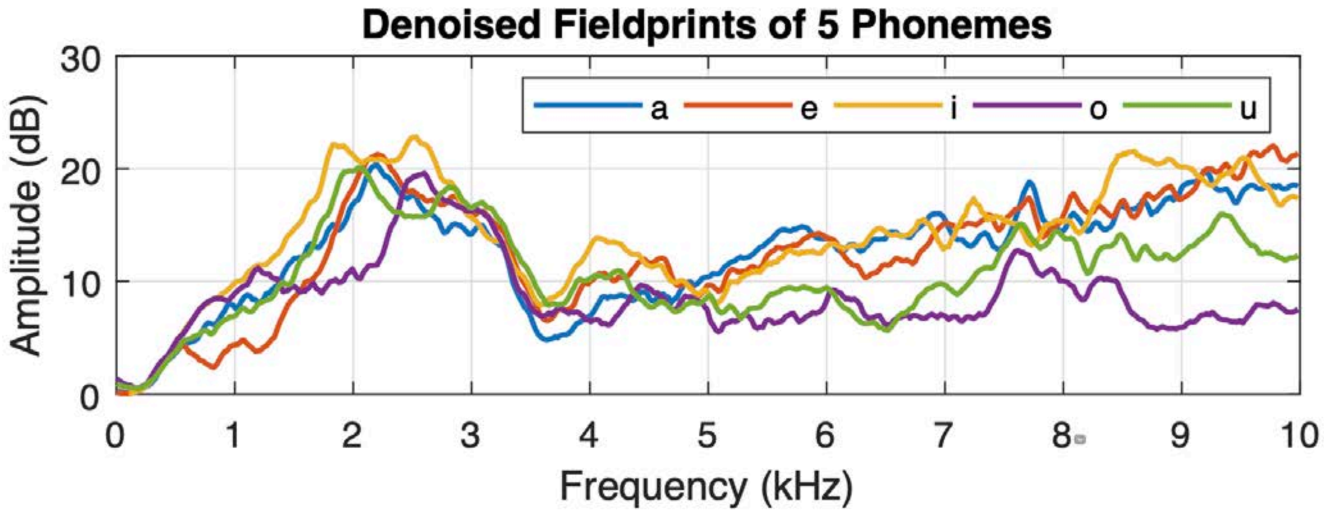
Q3 Fieldprint consistency – Speech content

❑ Challenge

Fieldprint changes with the speech content (phoneme)

❑ Solution: define LTAF

The human voice may approach a more phonetically balanced state for words and sentences



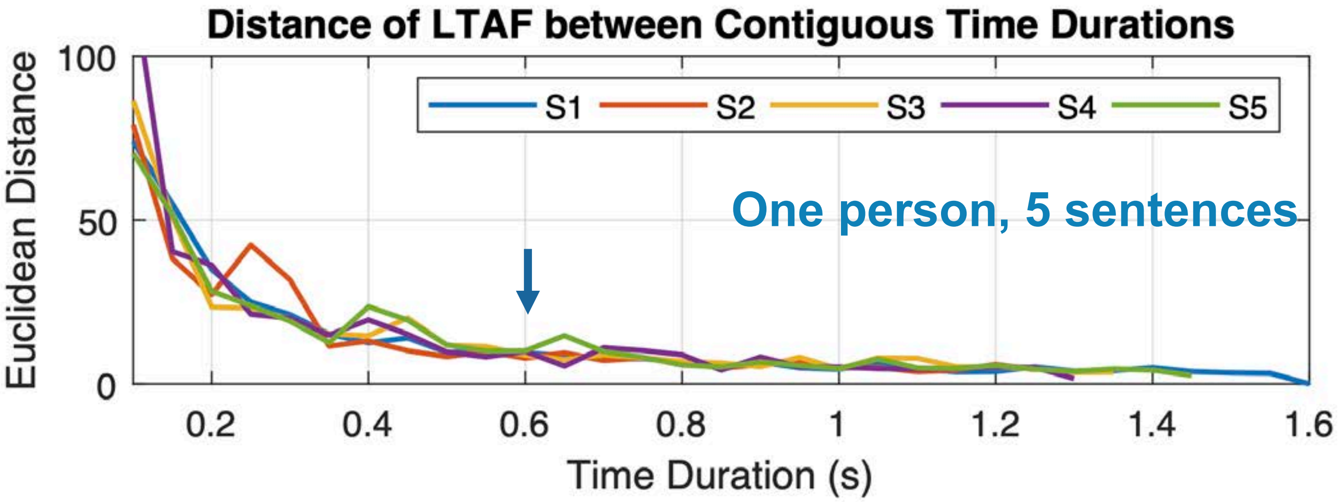
Long-Time Average Fieldprint (LTAF)

$$\mathcal{F}_{LTA}(\boldsymbol{p_1},\boldsymbol{p_2}) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{F}_i(\boldsymbol{p_1},\boldsymbol{p_2})$$

Q3 Fieldprint consistency – Speech content

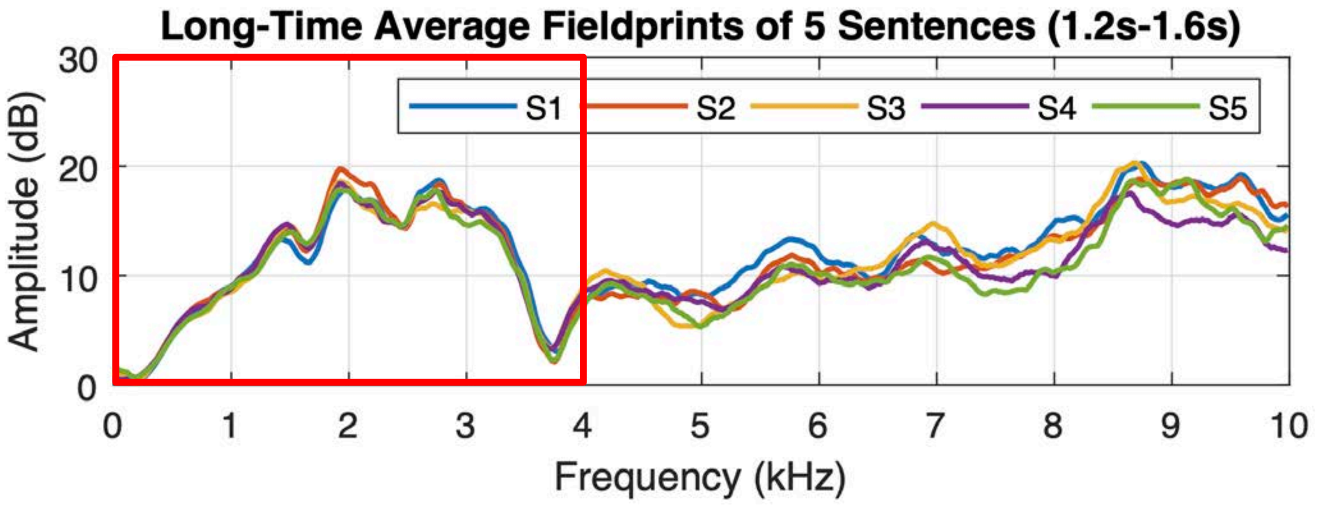
- Time duration

LTAF becomes more stable with a longer time duration



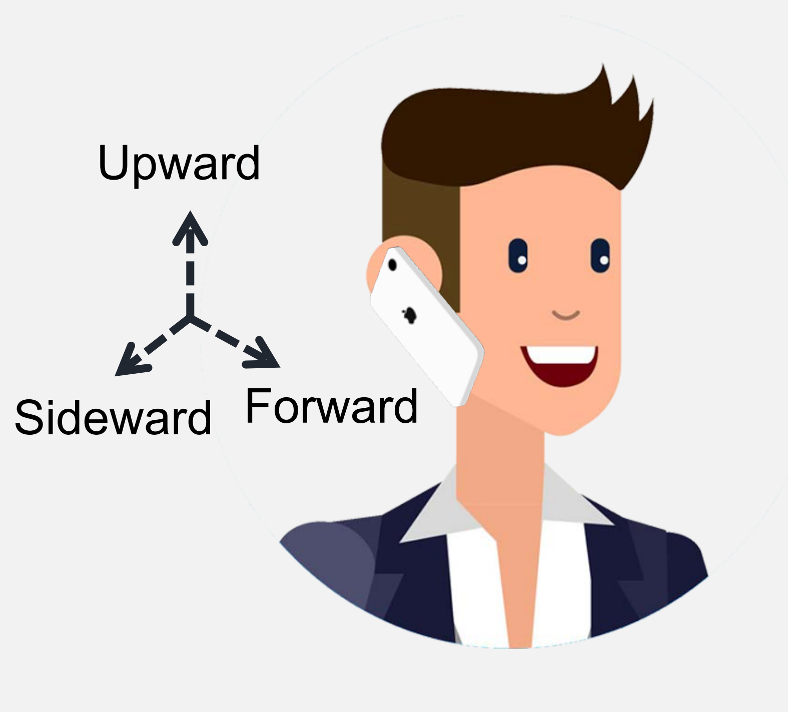
- Text-independent

The LTAFs of 5 different sentences are similar, especially below 4 kHz

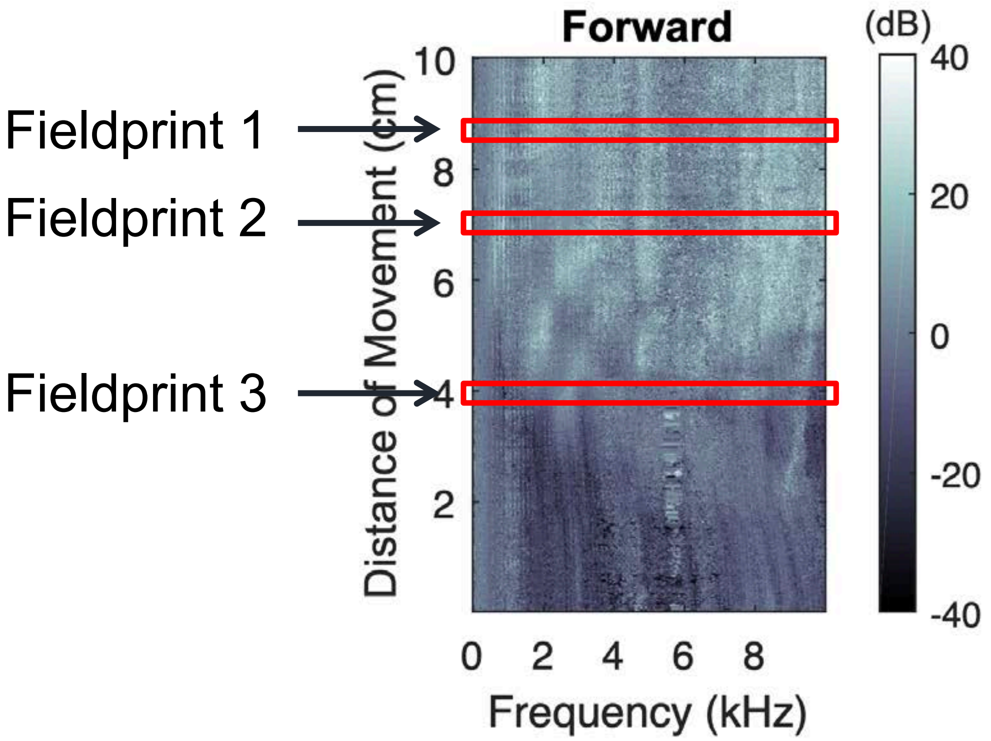


Q3 Fieldprint consistency – Microphone locations

Experiment setup



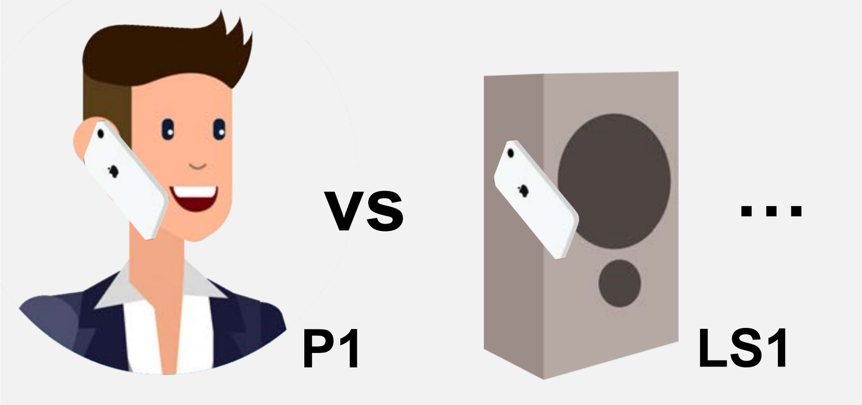
Fieldprints in colormap

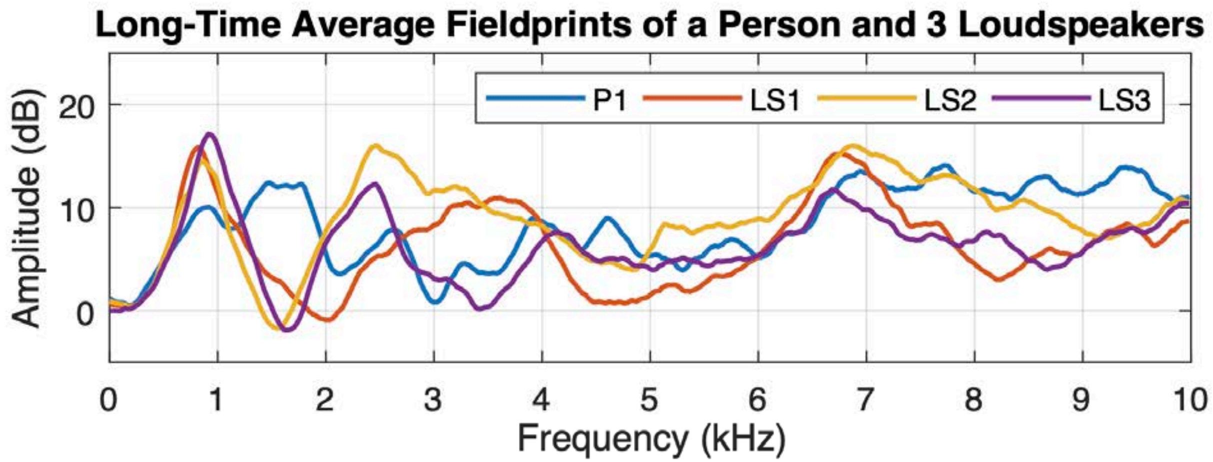


Fieldprint is consistent to modest microphone displacement

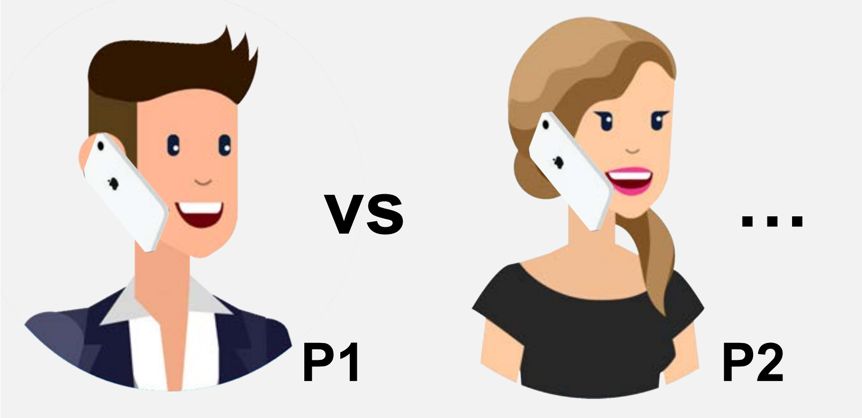
Q3 Fieldprint distinctiveness

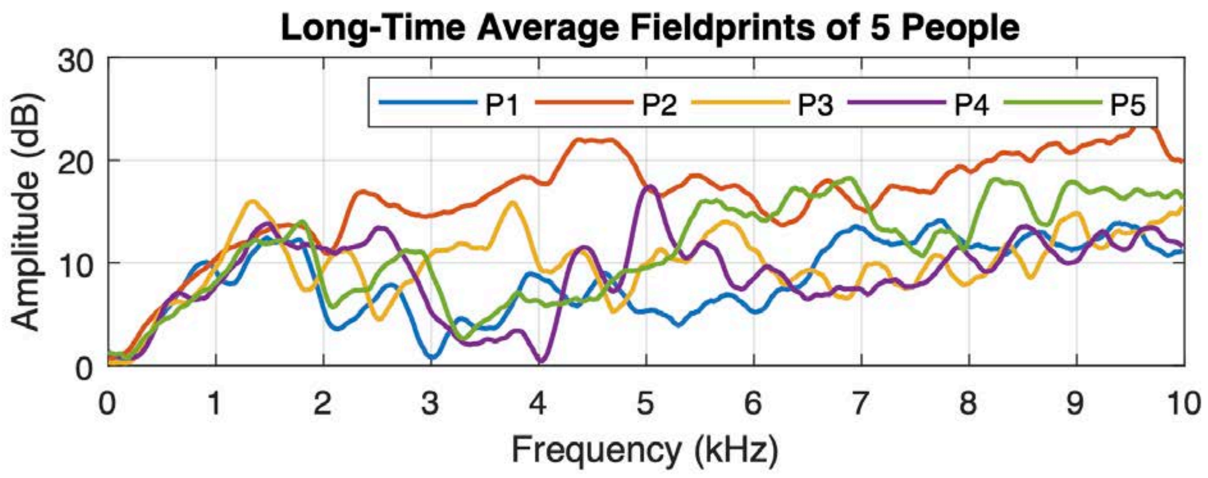
Between human and loudspeakers





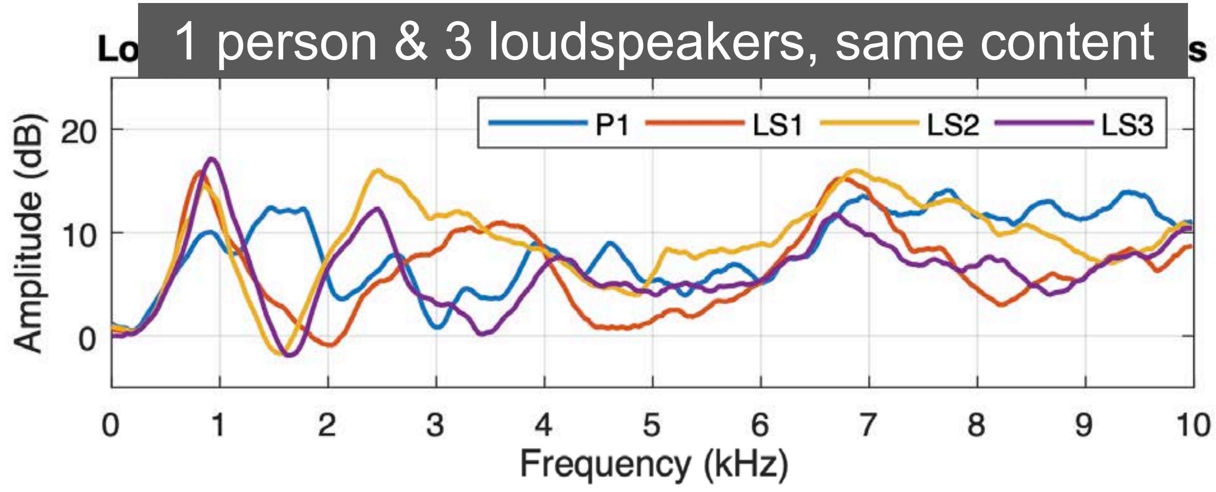
Between different people

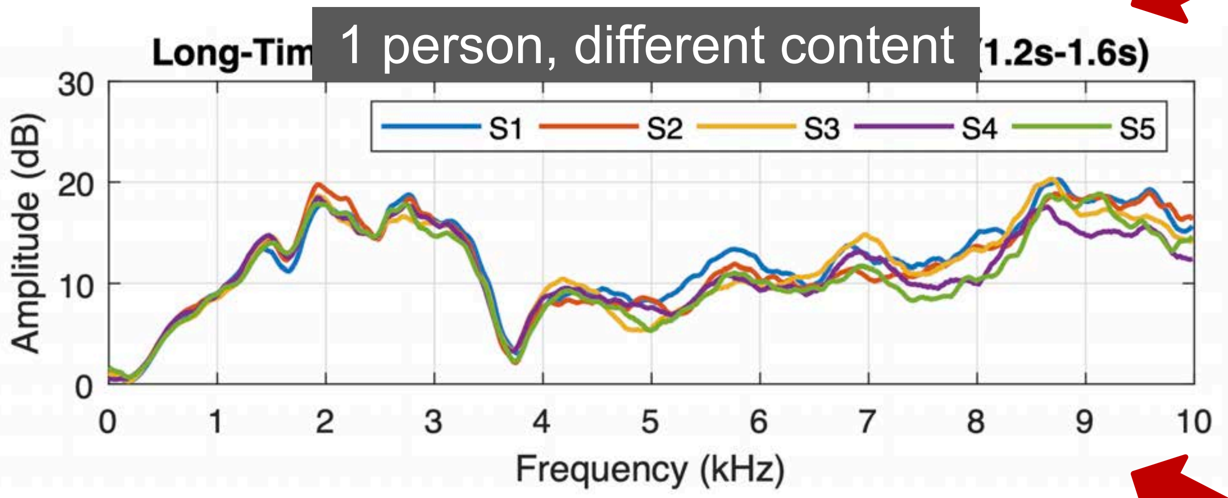


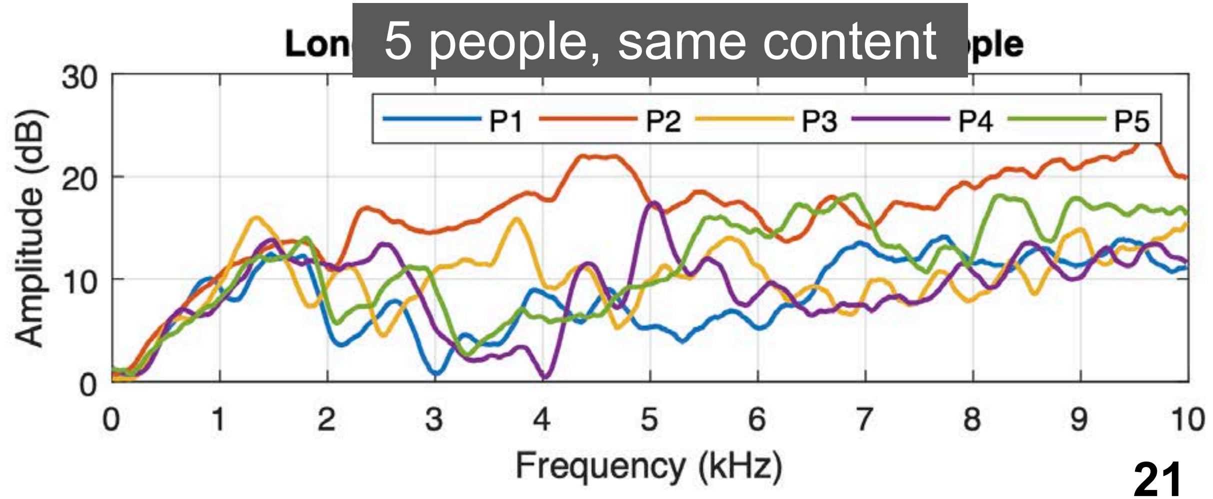


Q3 Fieldprint distinctiveness

Distinctive between a person and replaying loudspeakers







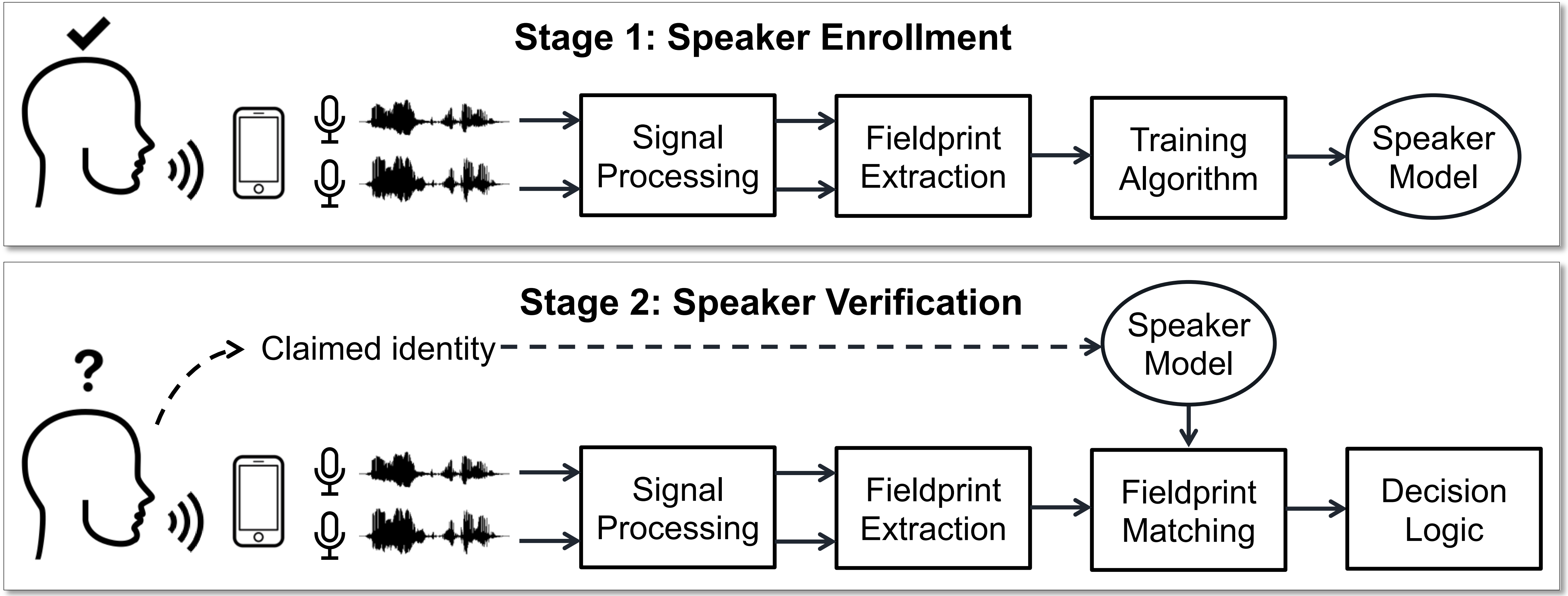
Distinctive between people

Fieldprint observations

- ❑ Consistency
  - Consistent as Long-Time Average Fieldprint (LTAF)
  - Text-independent
  - Consistent to modest microphone displacement
- ❑ Distinctiveness
  - Distinctive between human & loudspeakers and between people
- ❑ Usability
  - No extra device
  - User-defined device positions

Design - “The catcher in the (sound) field”

- **CaField**: a spoofing detection system based on fieldprints



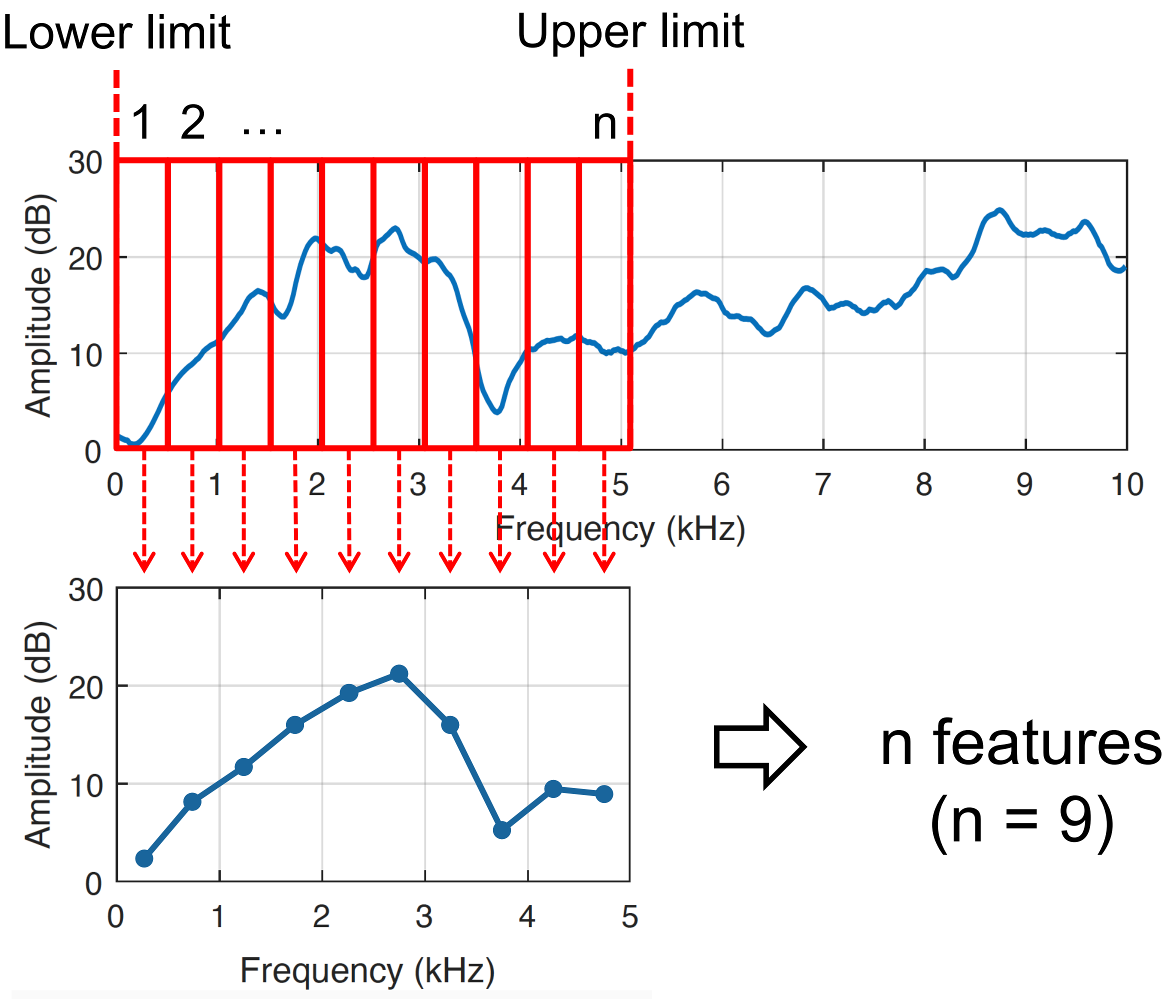
Design – Modules

❑ Fieldprint Extraction

- LTAF per command
- Low-dimensional features
- Filterbank (n bandpass filters)
- n dimensional feature vector

❑ Fieldprint matching

- ❑ Gaussian Mixture Model (GMM)
- ❑ Likelihood value
- ❑ Predefined threshold



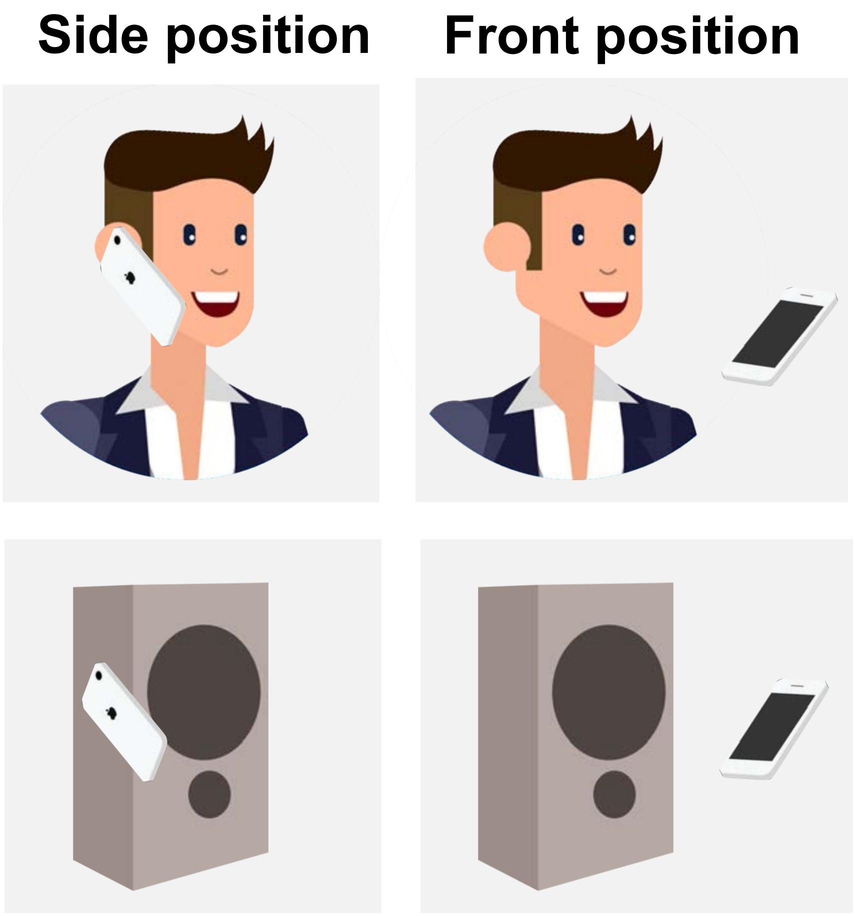
Evaluation – Dataset

❑ Human voice dataset

- 20 participants (6 female & 14 male)
- 2 types of device positions (side & front)
- Voice commands: 10 for enrollment & 40 for verification
- Total: 2,000 commands

❑ Spoofing attack (replay) dataset

- 8 loudspeakers of various sizes and qualities
- 2 types of device positions (side & front)
- Total: 16,000 spoofing commands



❑ Metrics

- Accuracy, Equal Error Rate (EER), False Acceptance Rate (FAR), False Rejection Rate (FRR)

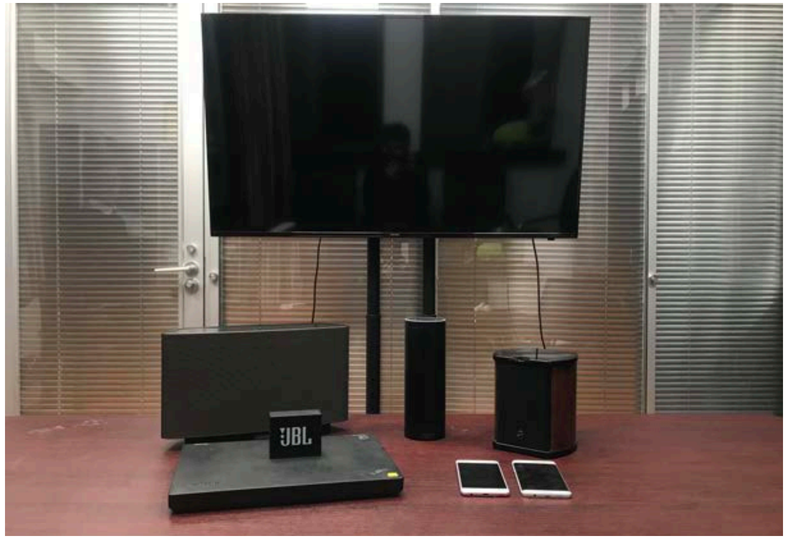
Evaluation – Dataset

❑ Human voice dataset

- 20 participants (6 female & 14 male)
- 2 types of device positions (side & front)
- Voice commands: 10 for enrollment & 40 for verification
- Total: 2,000 commands

❑ Spoofing attack (replay) dataset

- 8 loudspeakers of various sizes and qualities
- 2 types of device positions (side & front)
- Total: 16,000 spoofing commands



❑ Metrics

- Accuracy, Equal Error Rate (EER), False Acceptance Rate (FAR), False Rejection Rate (FRR)

Evaluation – Overall performance

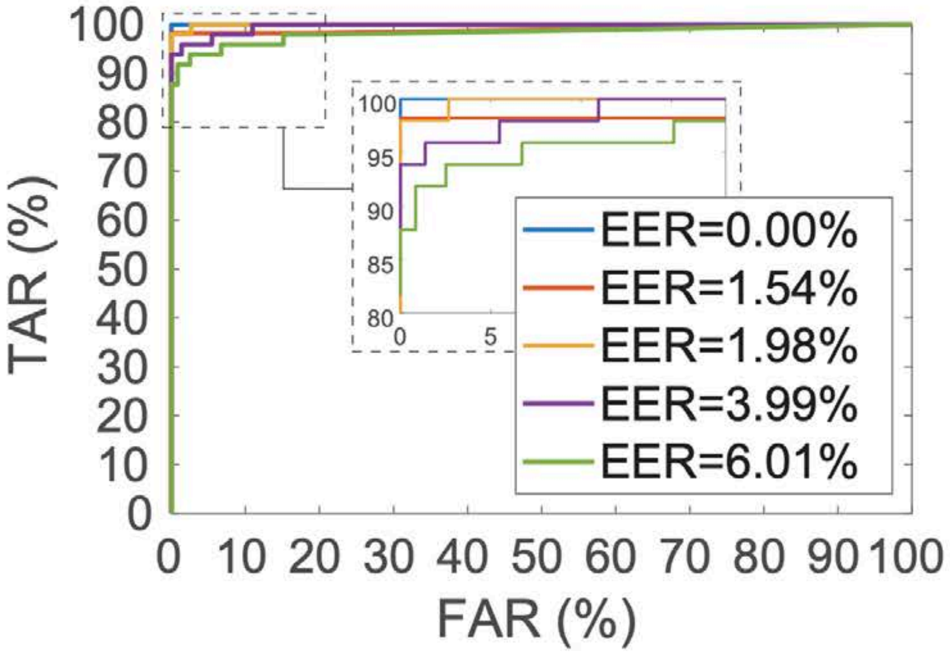
- ❑ Detecting spoofing attacks
- ❑ Differentiating human speakers

Function	Accuracy	FAR	FRR	EER
Detect spoofing attacks	99.16%	0.82%	0.97%	0.85%
Differentiate human speakers	98.42%	1.87%	1.43%	1.84%

CaField is highly effective in detecting spoofing attacks and differentiating different people

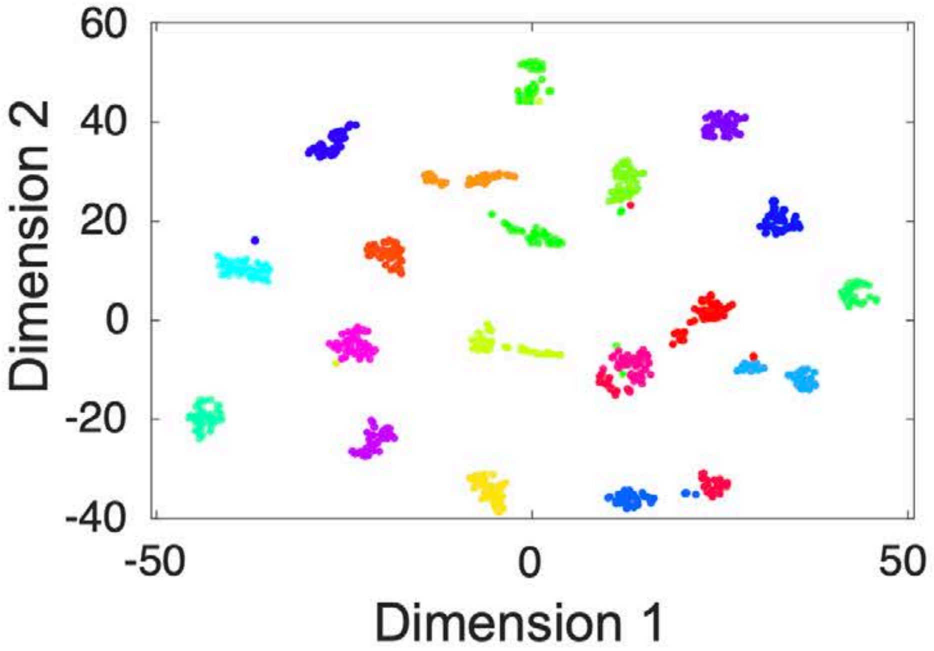
Evaluation – Overall performance

- ❑ Detecting spoofing attacks



ROC curves of 5 participants in spoofing detection

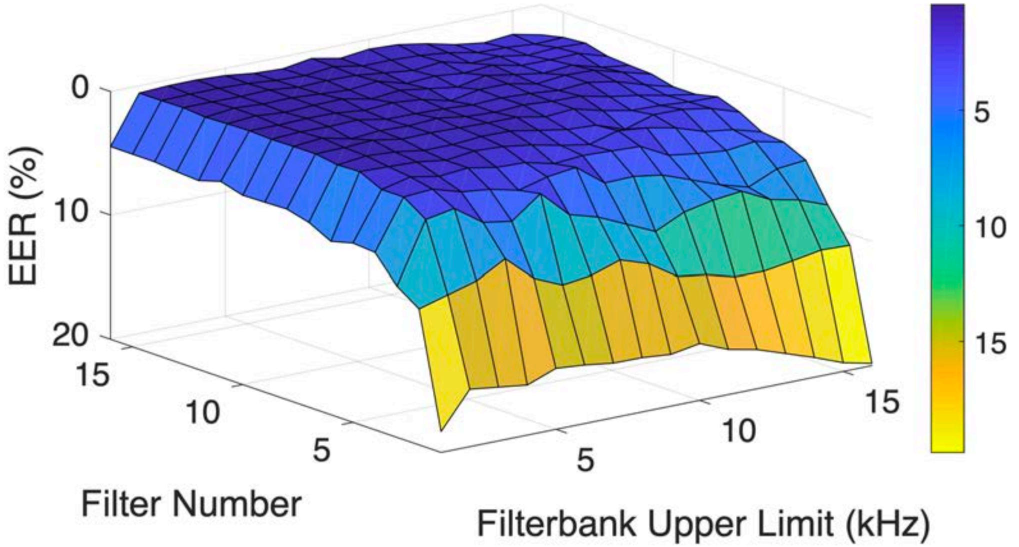
- ❑ Differentiating human speakers



Feature separation of 20 participants with t-SNE

Evaluation – Factors affecting spoofing detection

- System parameters
- Smartphone position
- Smartphone distance
- Type of loudspeaker
- Recording smartphone



More filters in the filterbank → higher performance
Freq. boundary > 5 kHz → performance slightly drops

CaField achieves a higher performance when the smartphone is on the side of the user

Impact of smartphone position

Position	Accuracy	FRR	FAR	EER
Front	98.74%	2.01%	1.16%	1.28%
Side	99.72%	0.63%	0.34%	0.38%

Conclusion

- ❑ Discovered the difference of **sound fields** between authentic users and spoofing attacks, and designed **fieldprint**
- ❑ Designed CaField, a fieldprint-based spoofing detection system
- ❑ Evaluation showed high performance in detecting attacks

- ❑ Future work
  - Arbitrary device positions across sessions
  - Replicating sound field with human-shaped loudspeakers