Crowd++: Unsupervised Speaker Count with Smartphones

ACM UbiComp'13

September 10th, 2013

Chenren Xu, Sugang Li, Gang Liu, Yanyong Zhang,
Emiliano Miluzzo, Yih-Farn Chen, Jun Li, Bernhard Firner

RUTGERS
WINLAB | Wireless Information Network Laboratory

at&t

UTD CRSS

Scenario 1: Dinner time, where to go?

Scenario 2: Is your kid social?

Scenario 2: Is your kid social?

Scenario 3: Which class is attractive?

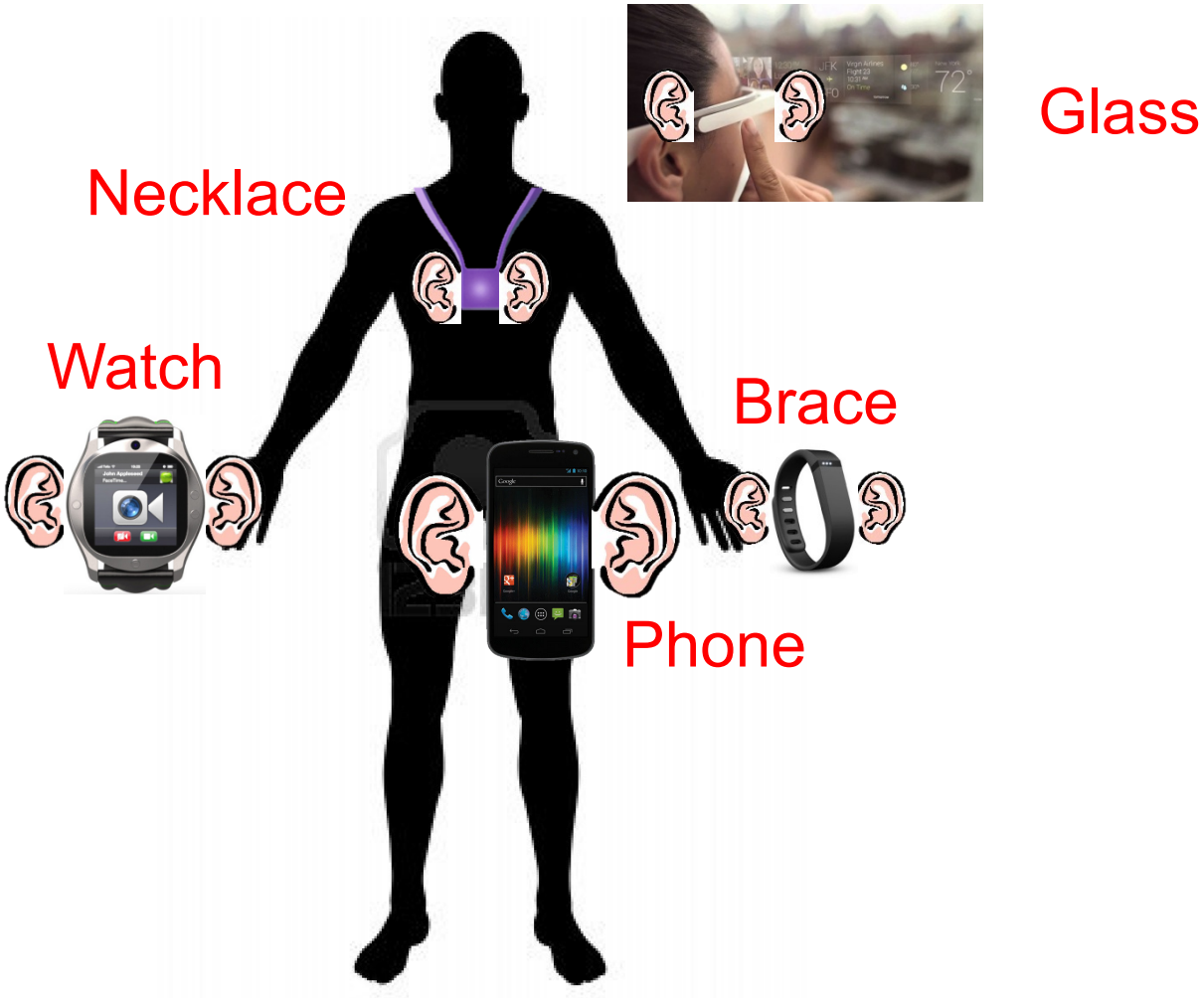
She will be my choice!

Solutions?

- Speaker count!
  - Dinner time, where to go?
    - Find the place where has most people talking!
  - Is your kid social?
    - Find how many (different) people they talked with!
  - Which class is more attractive?
    - Check how many students ask you questions!

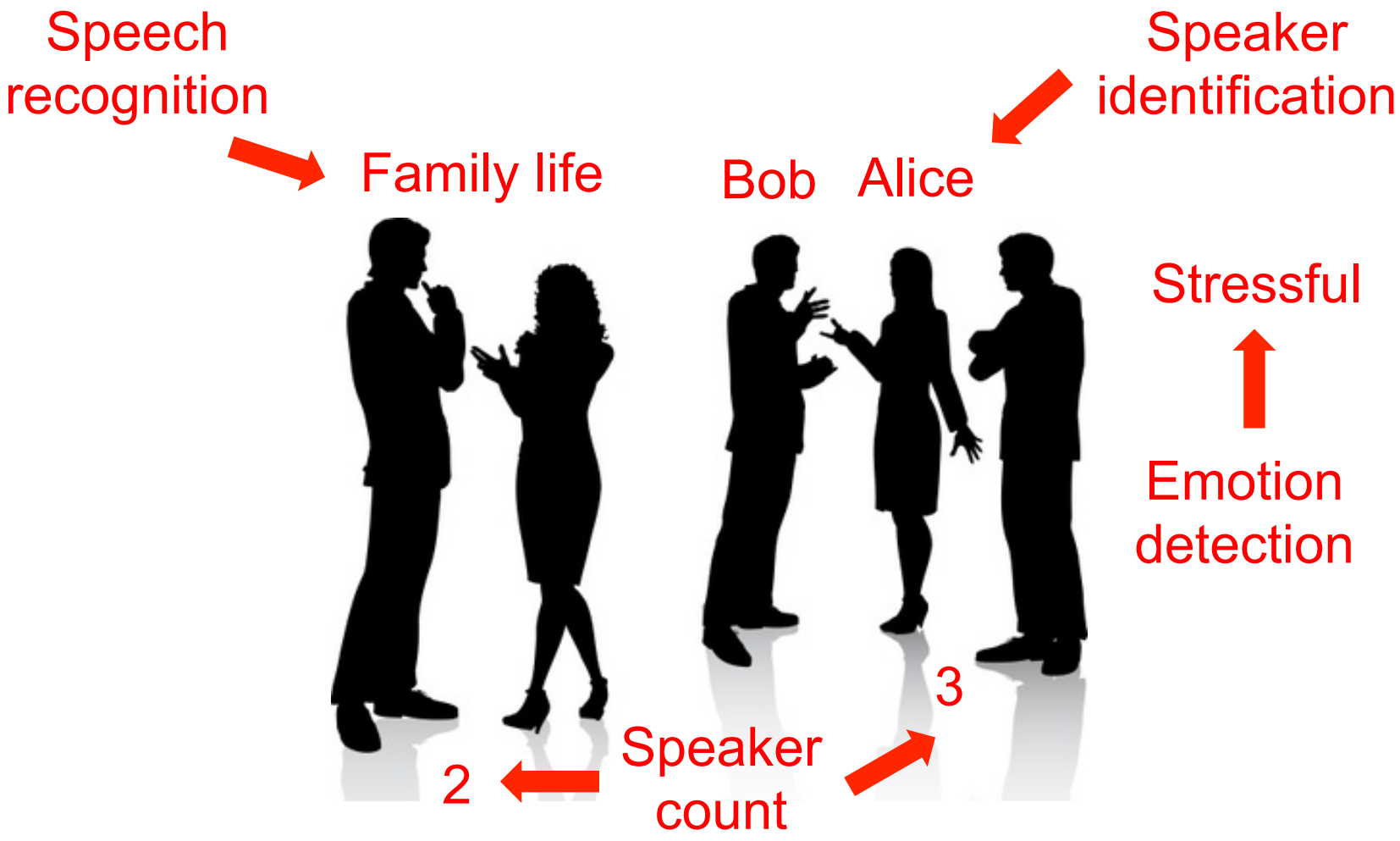
Microphone + microcomputer

The era of ubiquitous listening



What we already have

Voice-centric sensing



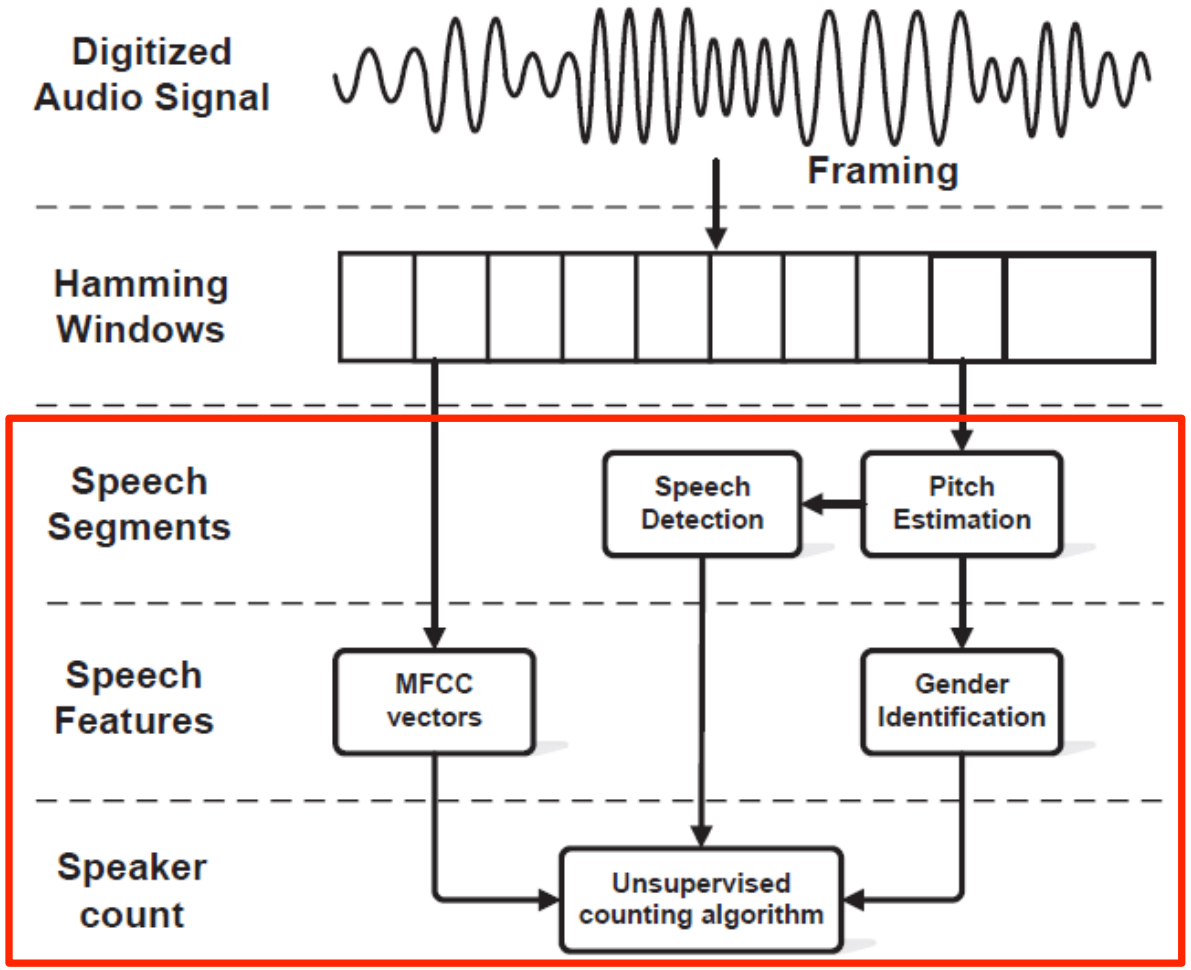
How to count?

Challenge
No prior knowledge of speakers
Background noise
Speech overlap
Energy efficiency
Privacy concern

How to count?

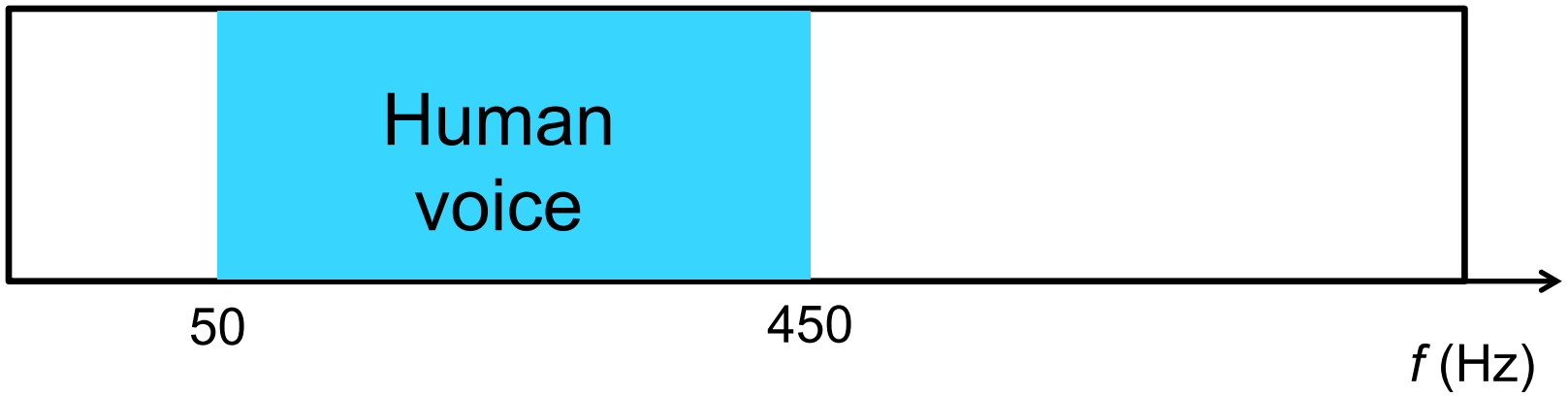
Challenge	Solution
No prior knowledge of speakers	Unique features extraction
Background noise	Frequency-based filter
Speech overlap	Overlap detection
Energy efficiency	Coarse-grained modeling
Privacy concern	On-device computation

Overview of Crowd++



Speech detection

- Pitch-based filter
  - Determined by the vibratory frequency of the vocal folds
  - Human voice statistics: spans from 50 Hz to 450 Hz



Speaker features

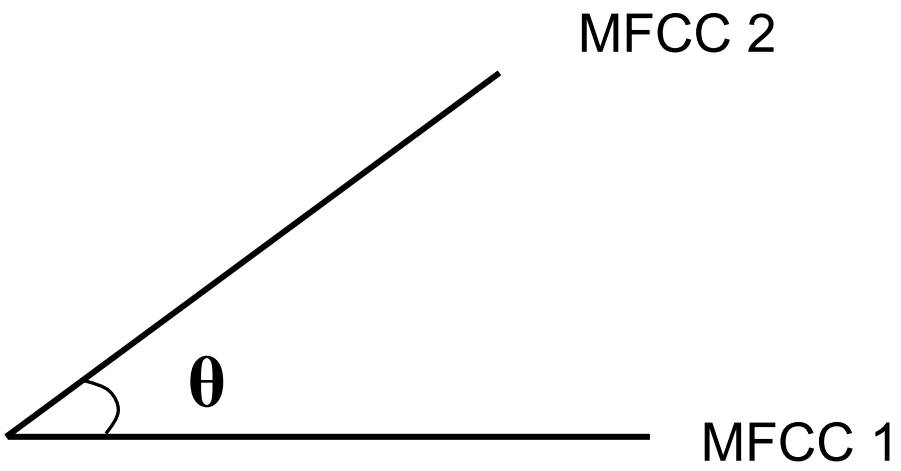
- MFCC
  - Speaker identification/verification
    - Alice or Bob, or else?
  - Emotion/stress sensing
    - Happy, or sad, stressful, or fear, or anger?
  - Speaker counting
    - No prior information

Supervised

Unsupervised

Speaker features

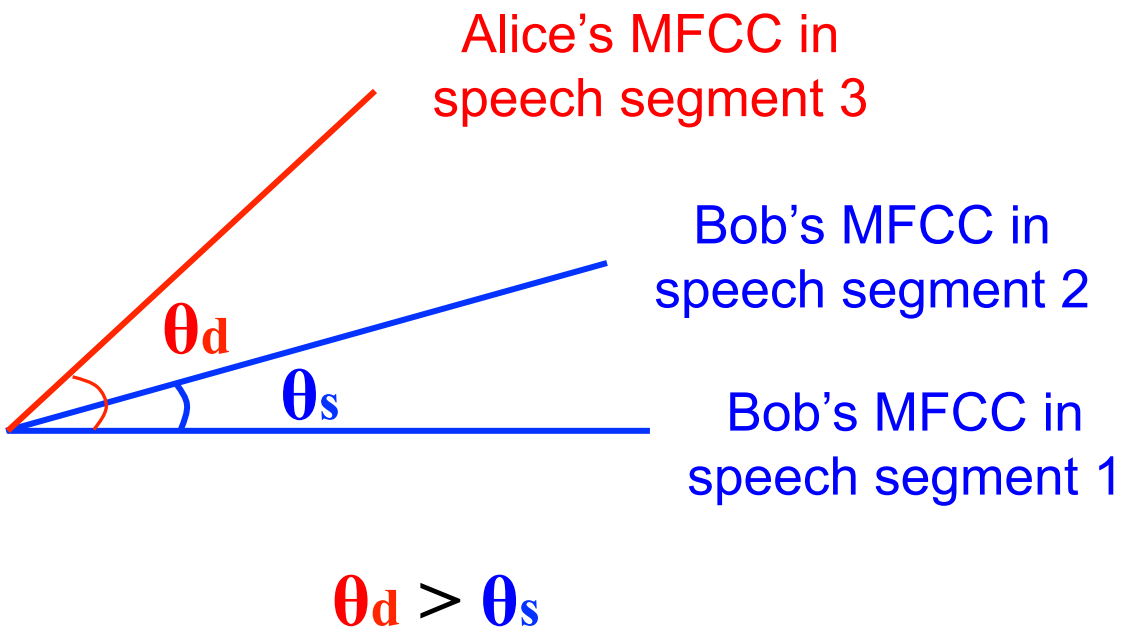
- Same speaker or not?
  - MFCC + cosine similarity distance metric



We use the angle θ **to capture the distance between speech segments.**

Speaker features

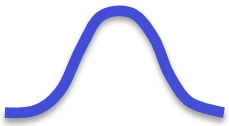
- MFCC + cosine similarity distance metric



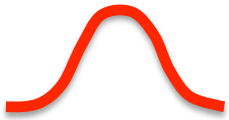
$\theta_d > \theta_s$

Speaker features

- MFCC + cosine similarity distance metric

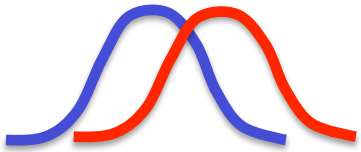


histogram of θ_s

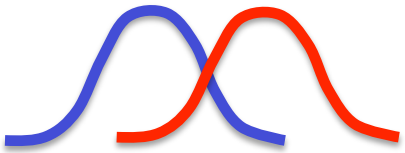


histogram of θ_d

1 second
speech segment



2-second
speech segment



3-sec
speech s

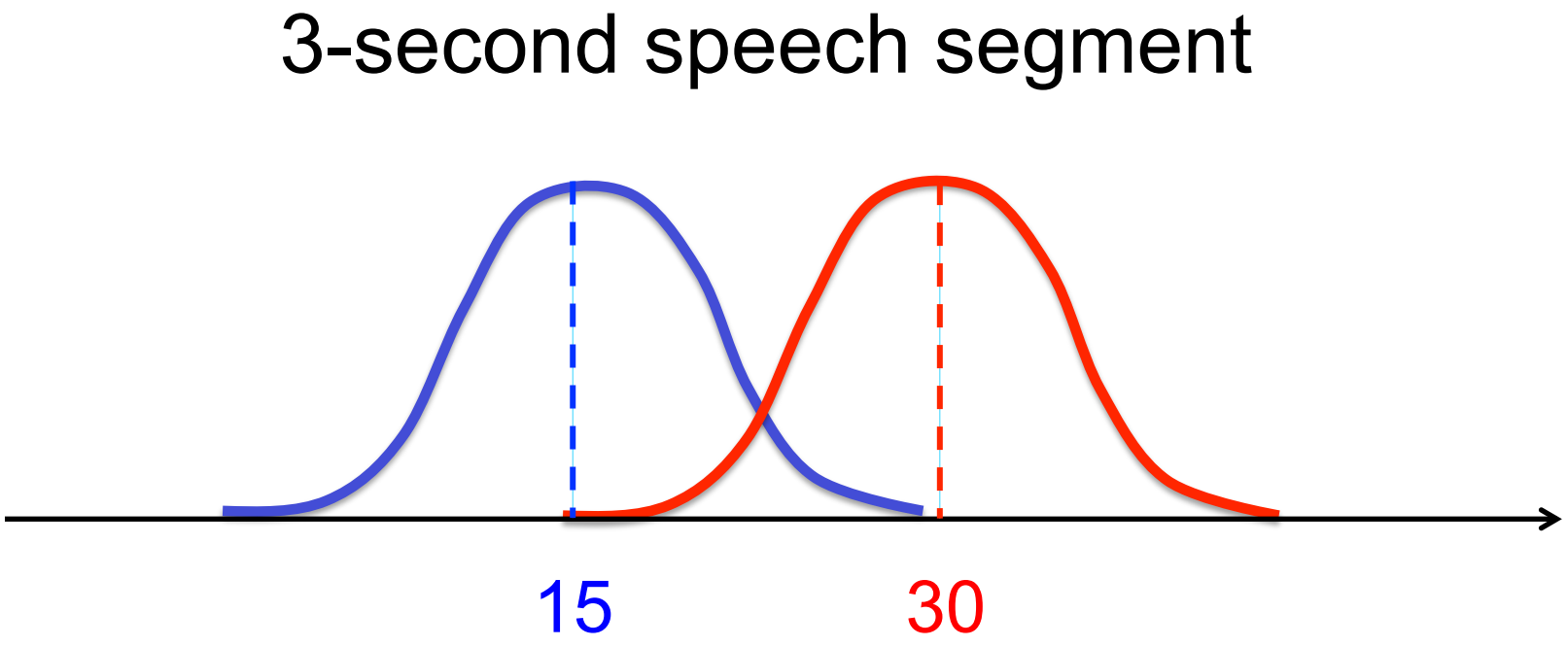
10-sec
speech s

We use 3-second for basic speech unit.

10 seconds is not natural in conversation!

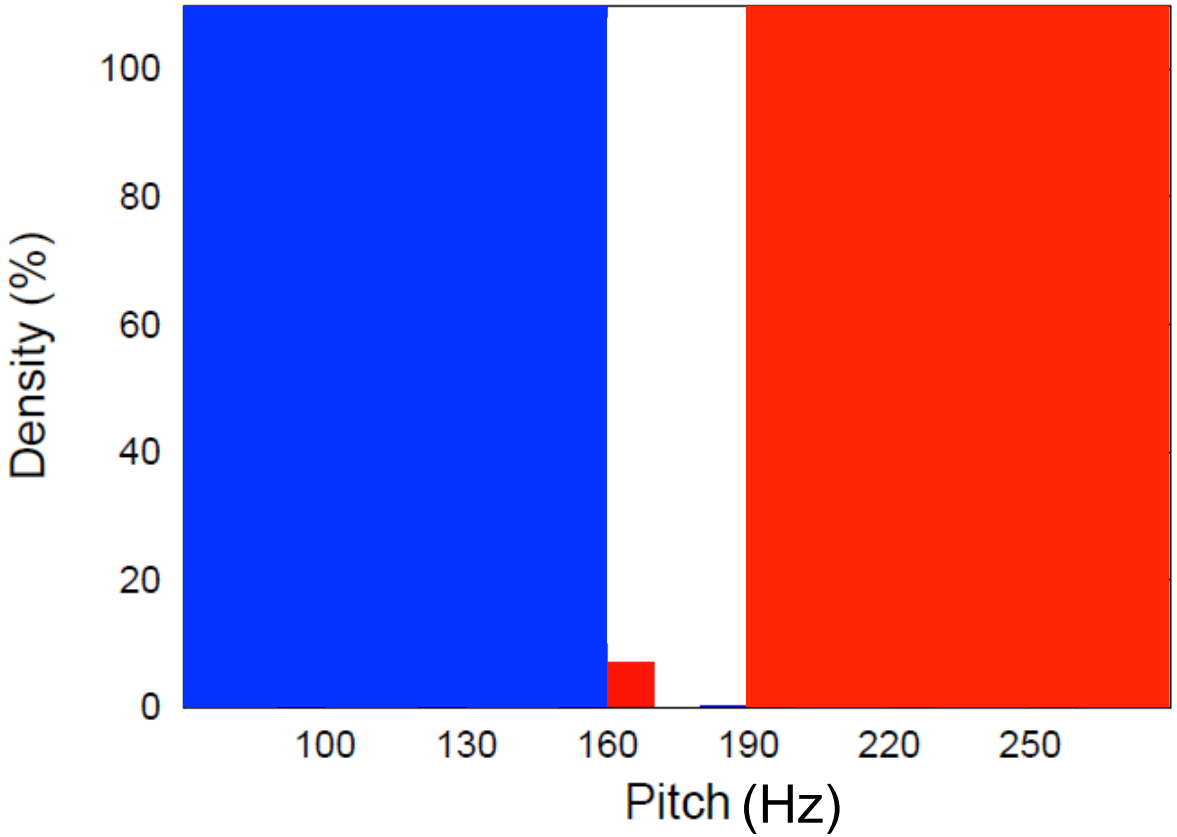
Speaker features

- MFCC + cosine similarity distance metric



Speaker features

- Pitch + gender statistics



Same speaker or not?

IF MFCC cosine similarity score < 15

AND

Pitch indicates they are same gender

Same speaker

ELSEIF MFCC cosine similarity score > 30

OR

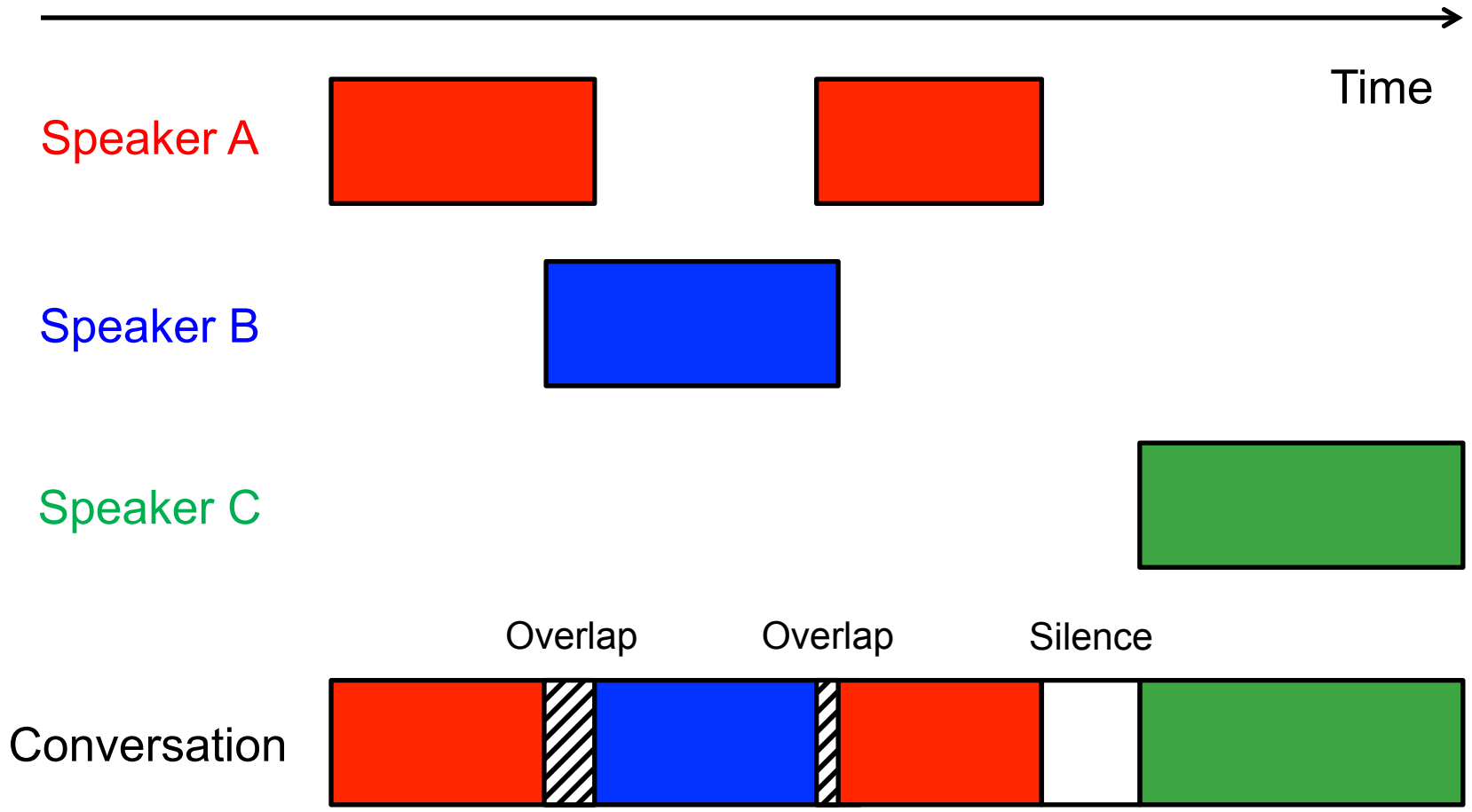
Pitch indicates they are different genders

Different speakers

ELSE

Not sure

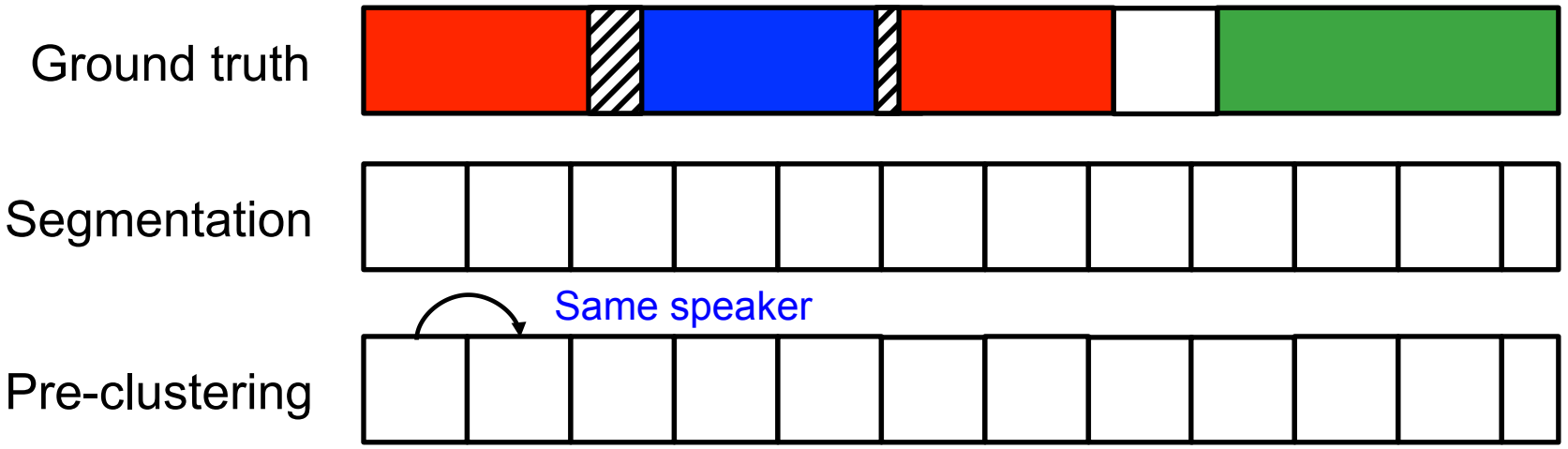
Conversation example



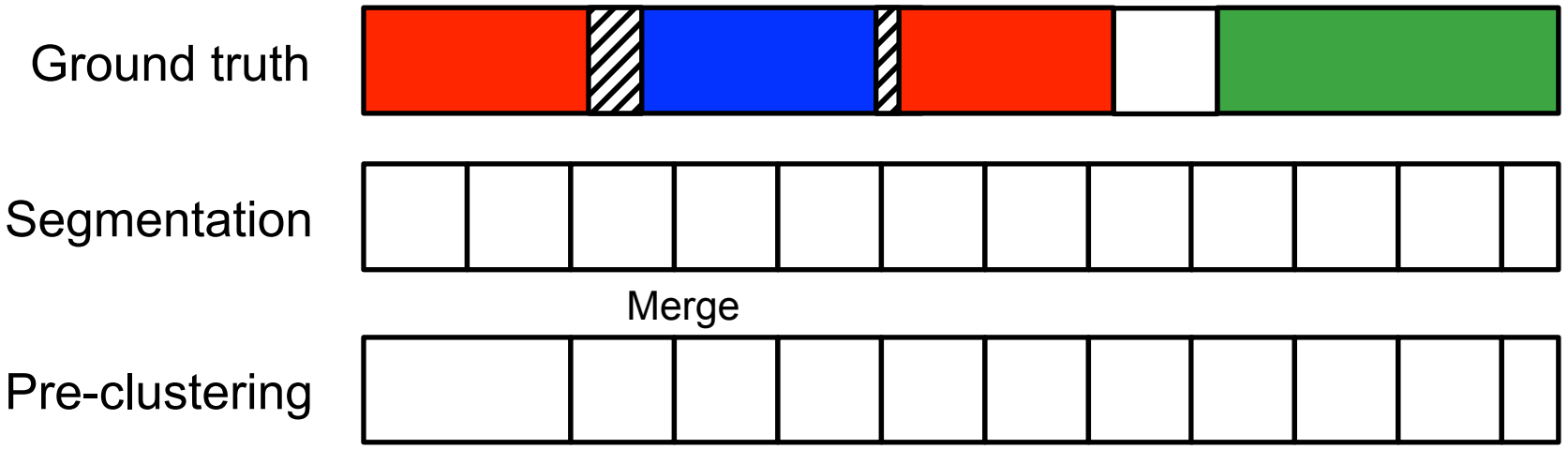
Unsupervised speaker counting

- Phase 1: pre-clustering
  - Merge the speech segments from same speakers

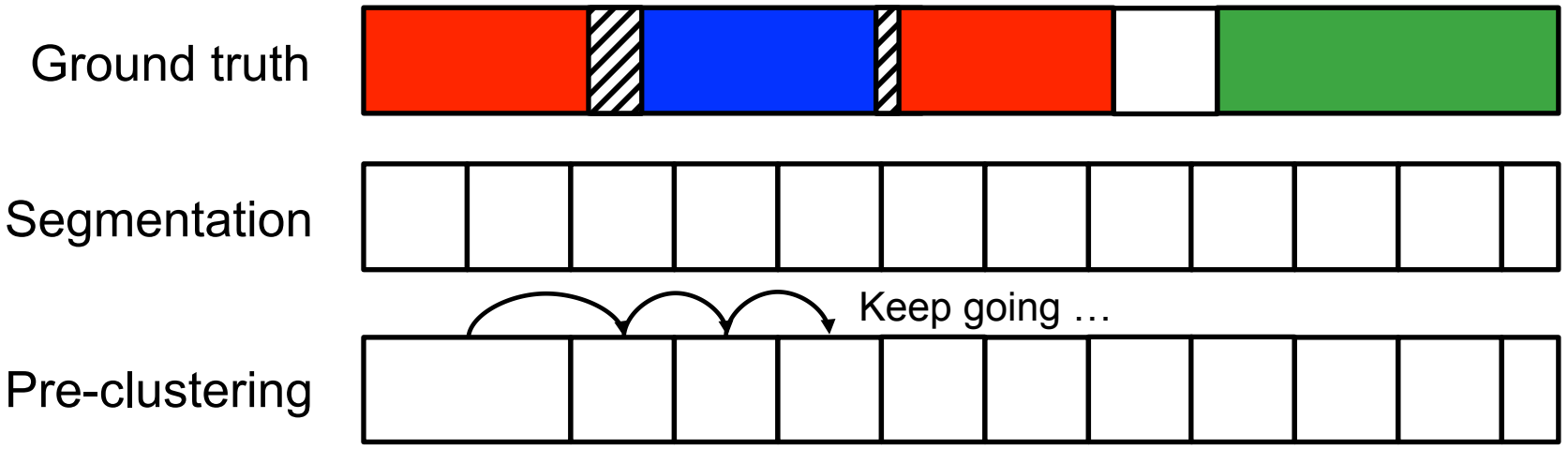
Unsupervised speaker counting



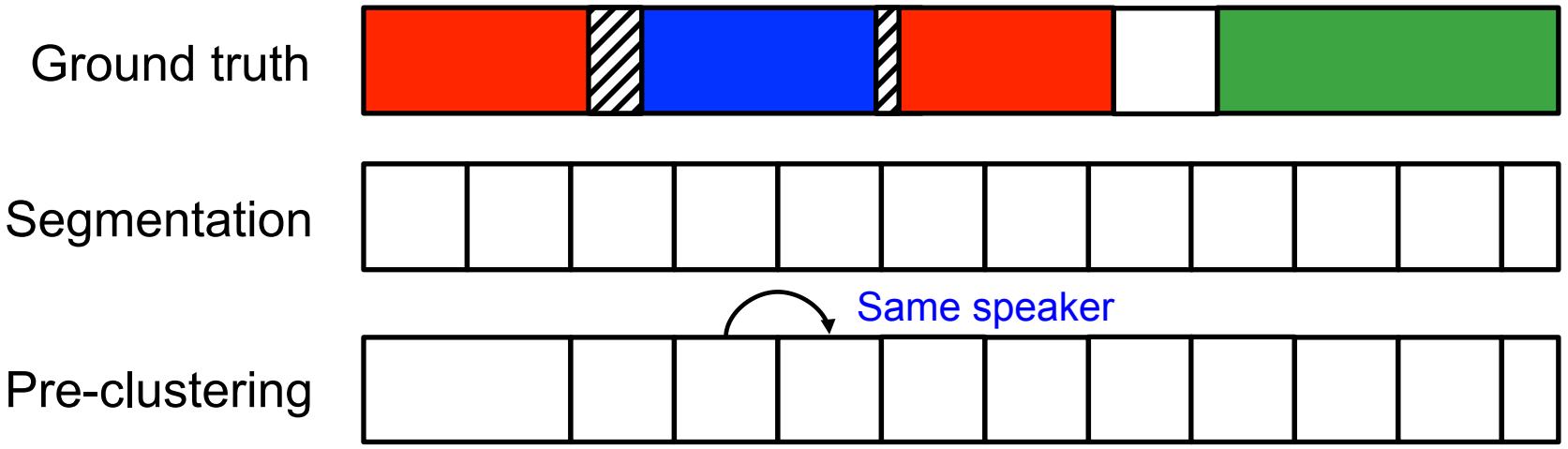
Unsupervised speaker counting



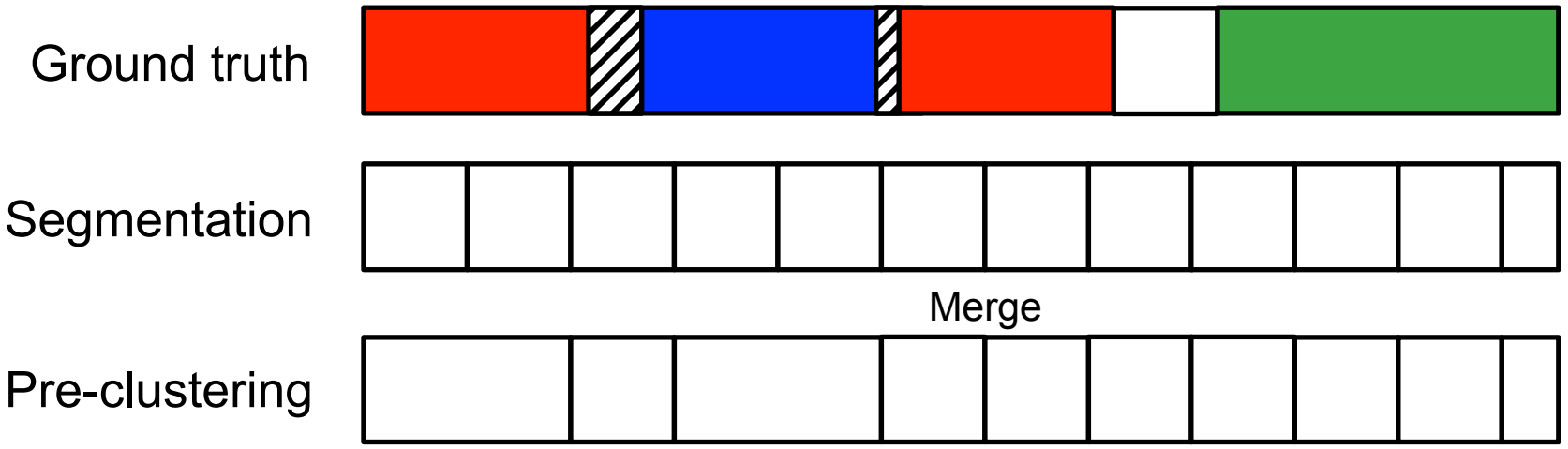
Unsupervised speaker counting



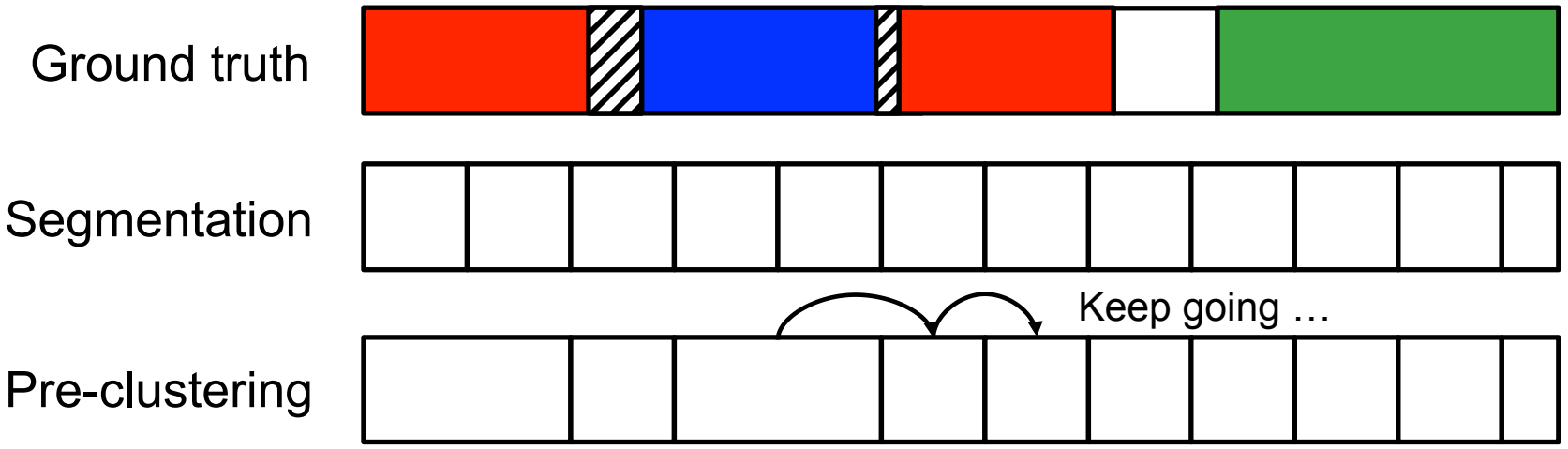
Unsupervised speaker counting



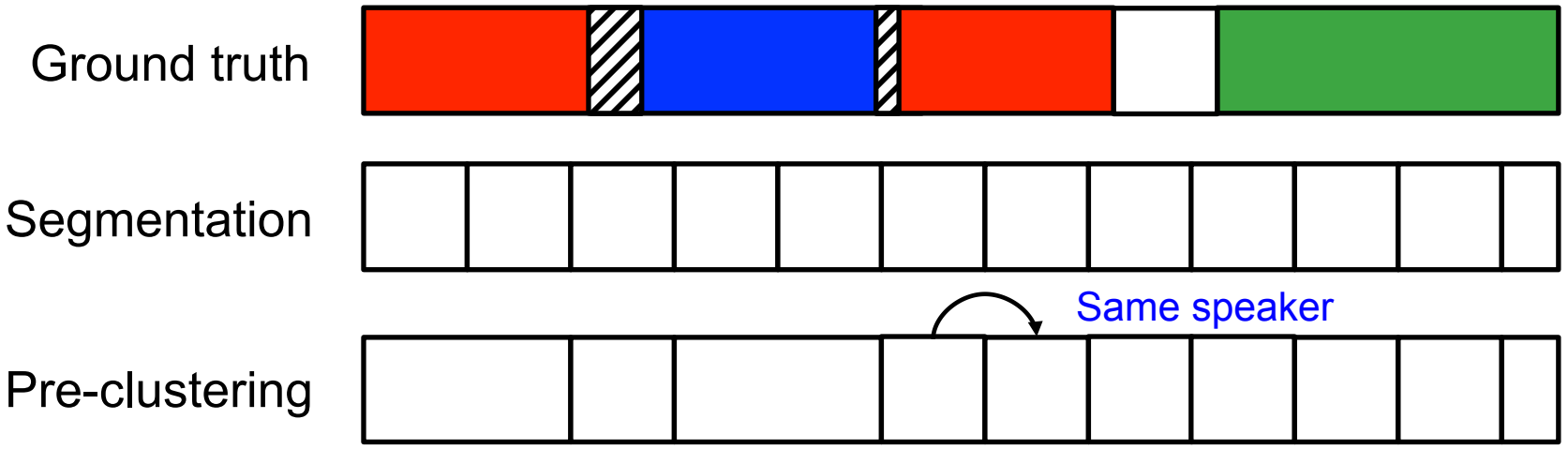
Unsupervised speaker counting



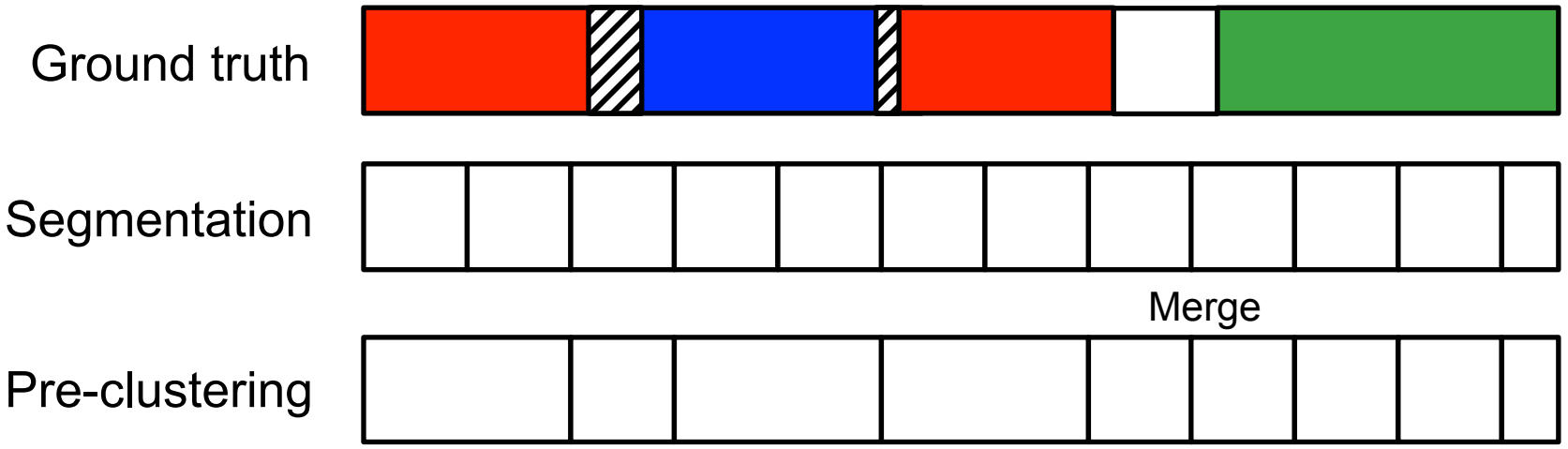
Unsupervised speaker counting



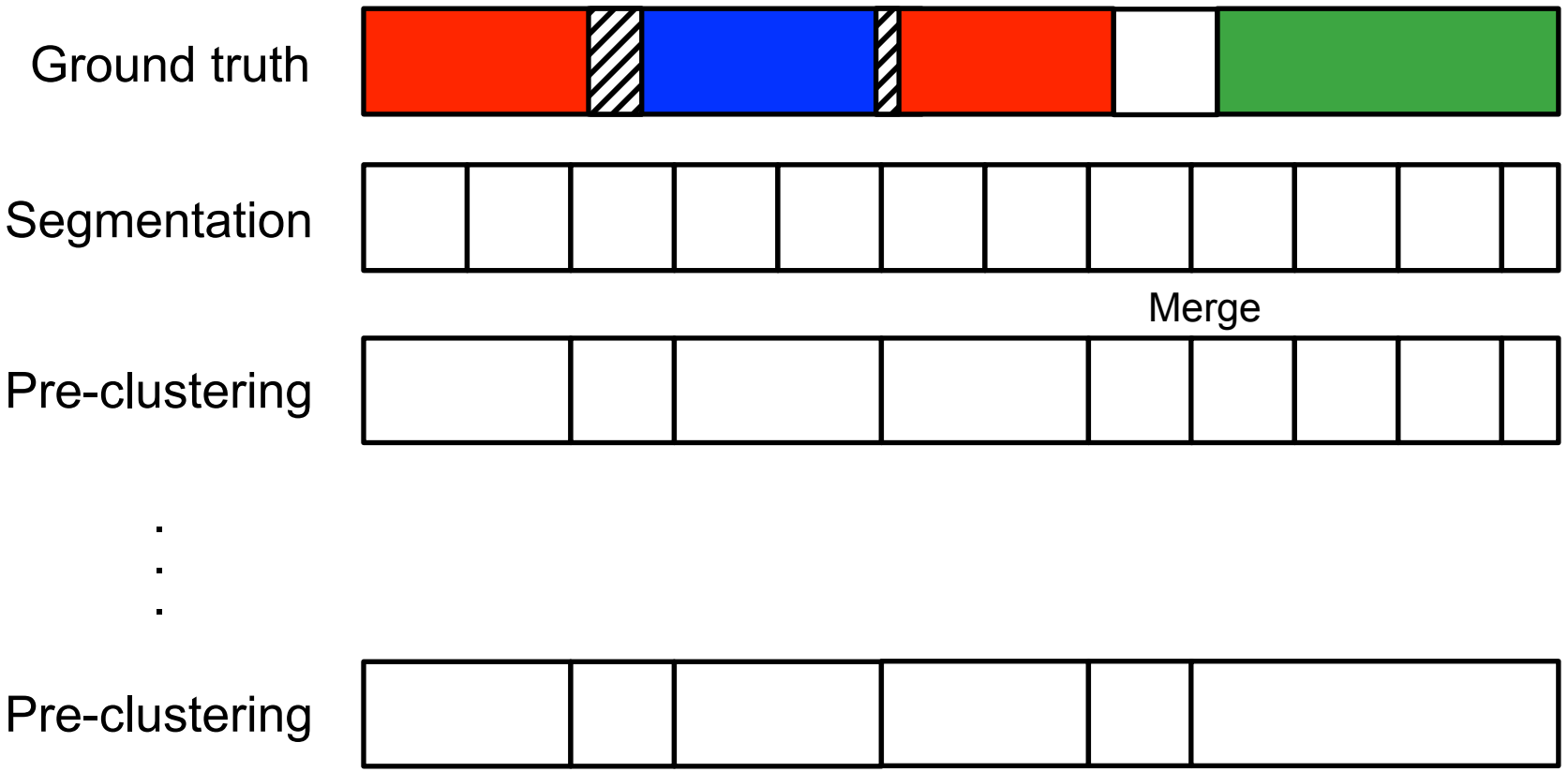
Unsupervised speaker counting



Unsupervised speaker counting



Unsupervised speaker counting



Unsupervised speaker counting

- Phase 1: pre-clustering
  - Merge the speech segments from same speakers
- Phase 2: counting
  - Only admit new speaker when its speech segment is different from all the admitted speakers.
  - Dropping uncertain speech segments.

Unsupervised speaker counting

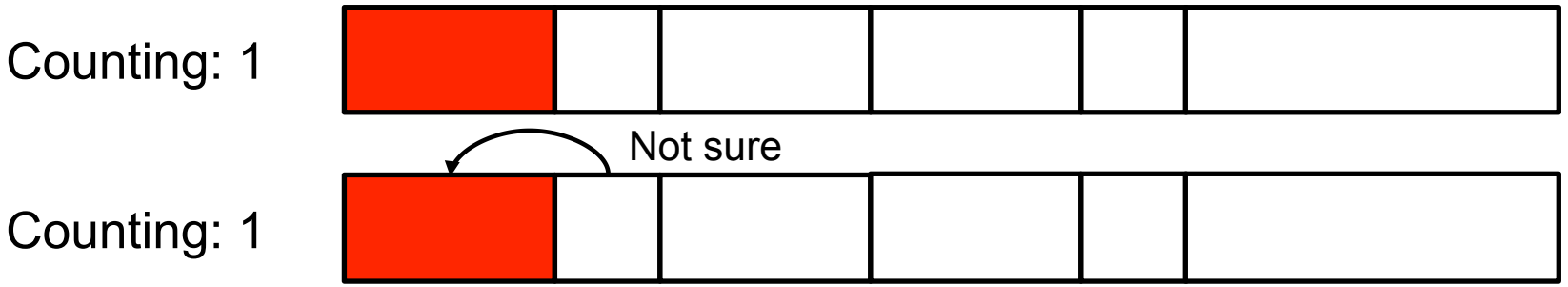
Counting: 0

Unsupervised speaker counting

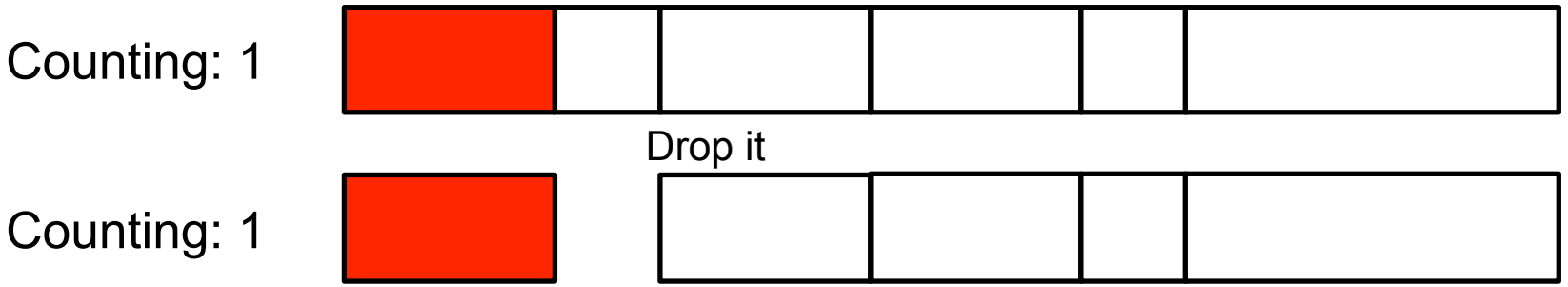
Counting: 1

Admit first speaker

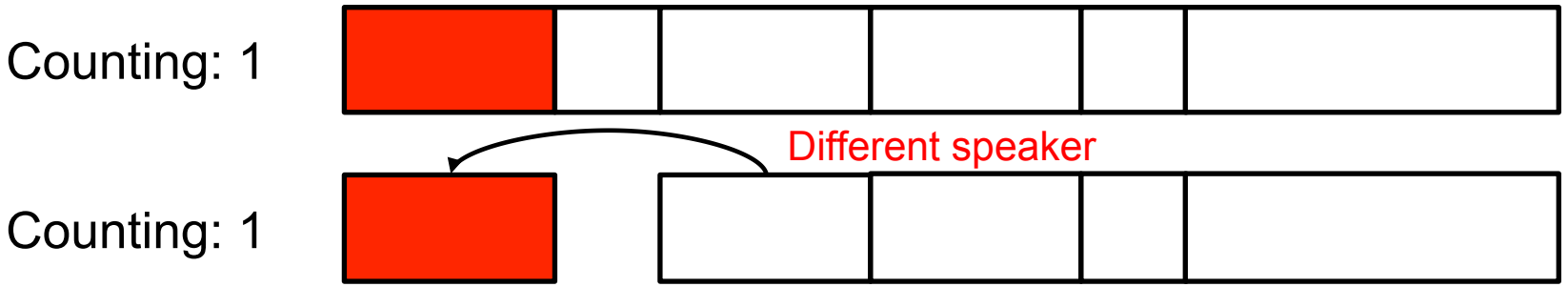
Unsupervised speaker counting



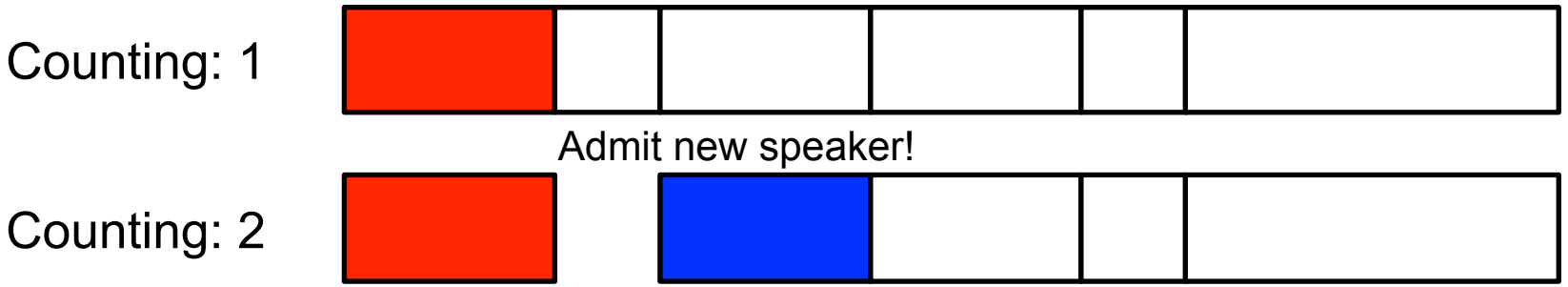
Unsupervised speaker counting



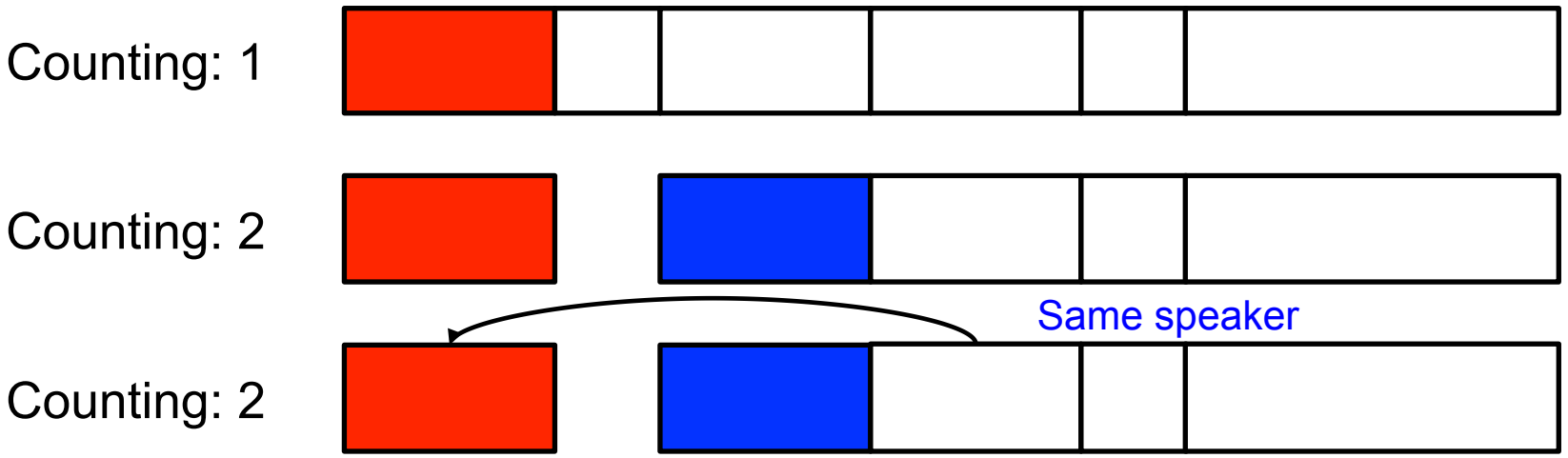
Unsupervised speaker counting



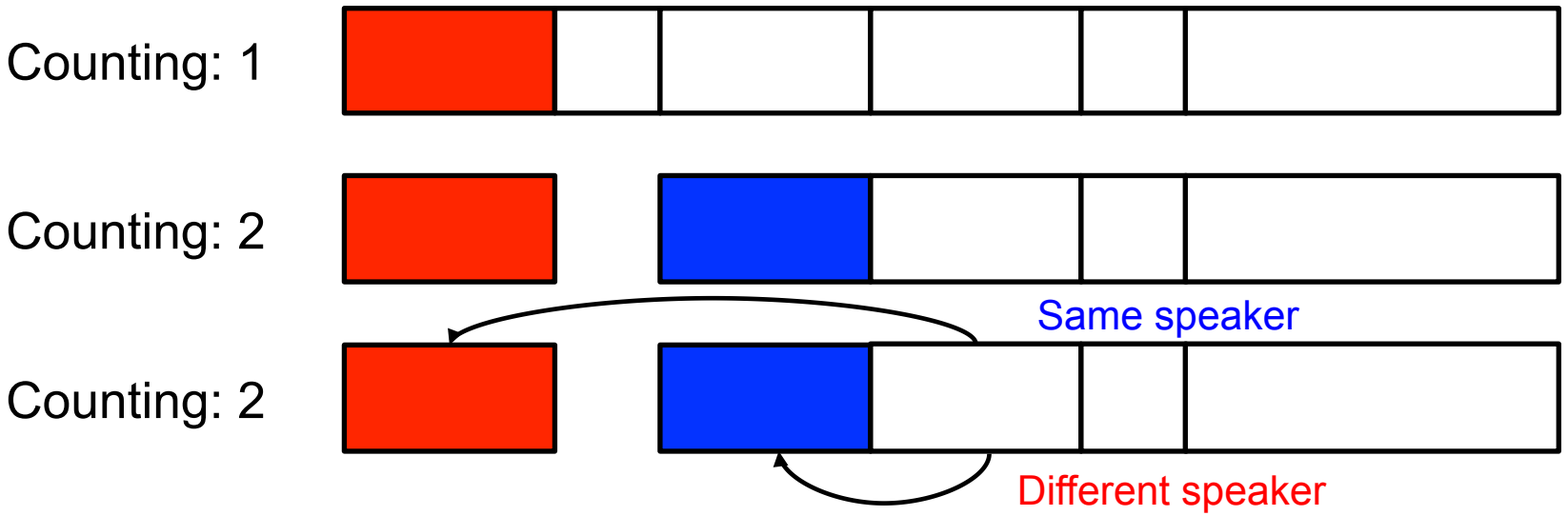
Unsupervised speaker counting



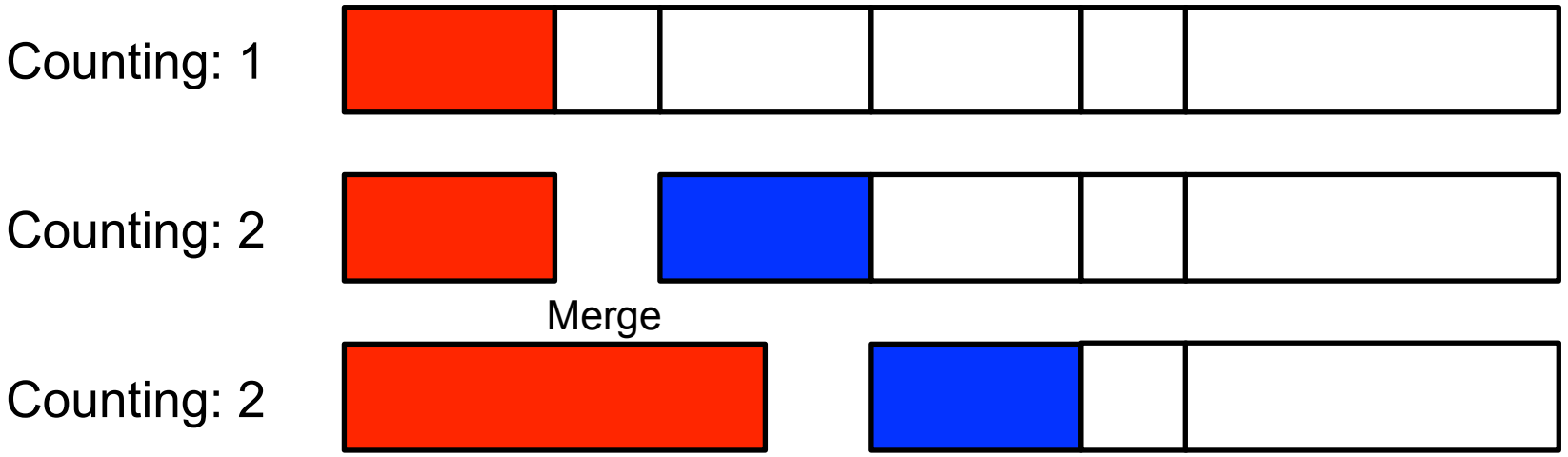
Unsupervised speaker counting



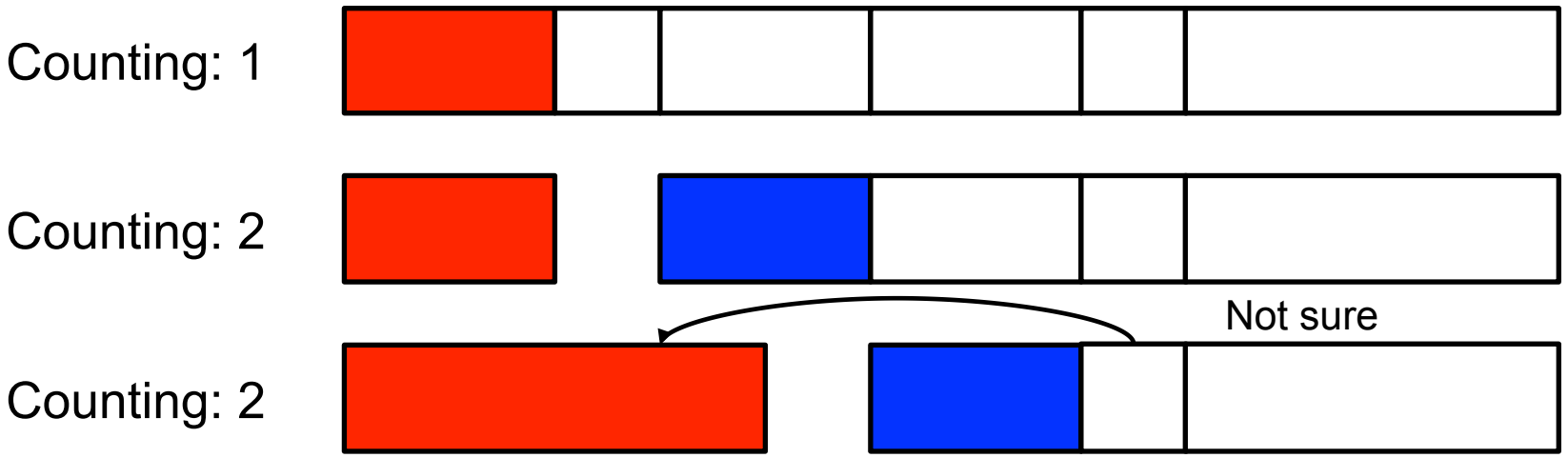
Unsupervised speaker counting



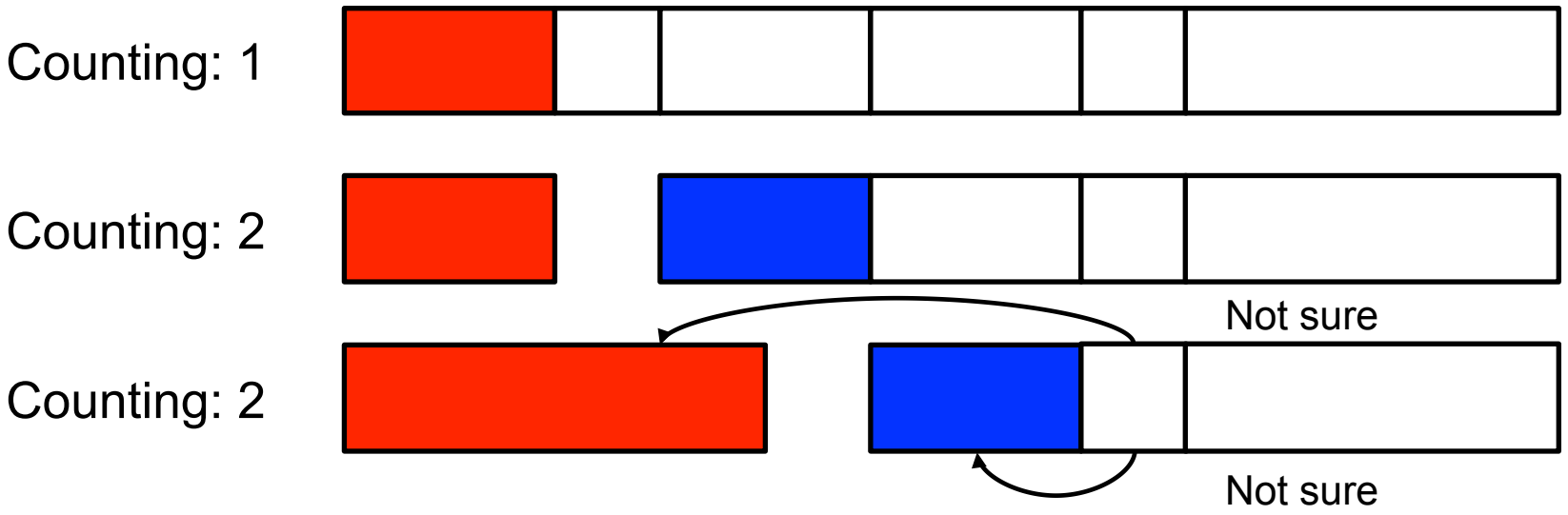
Unsupervised speaker counting



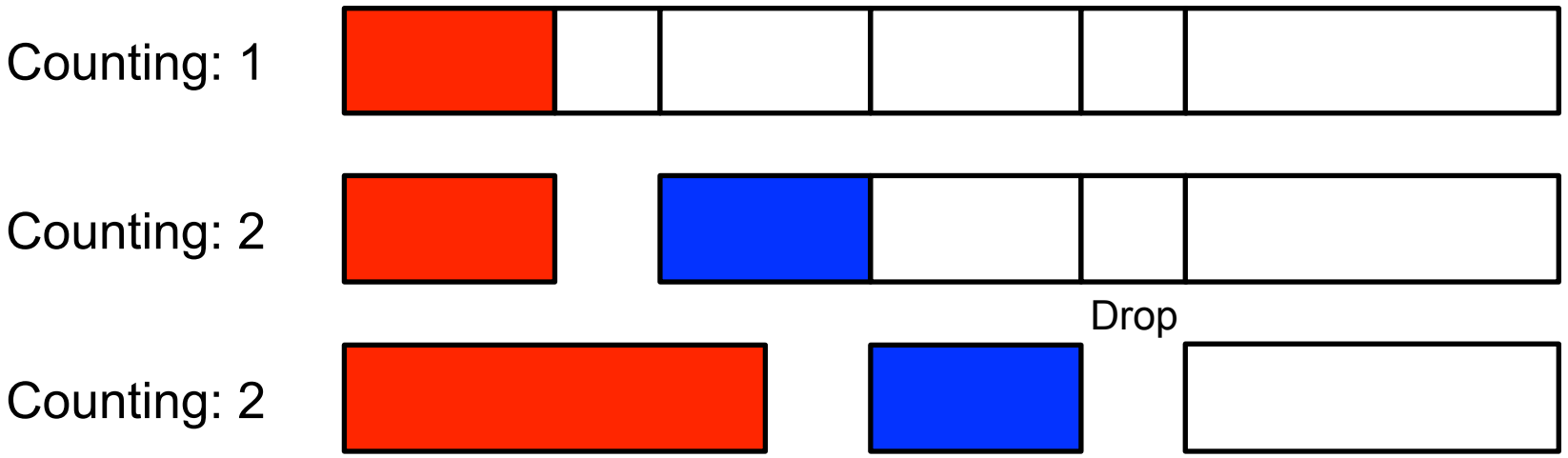
Unsupervised speaker counting



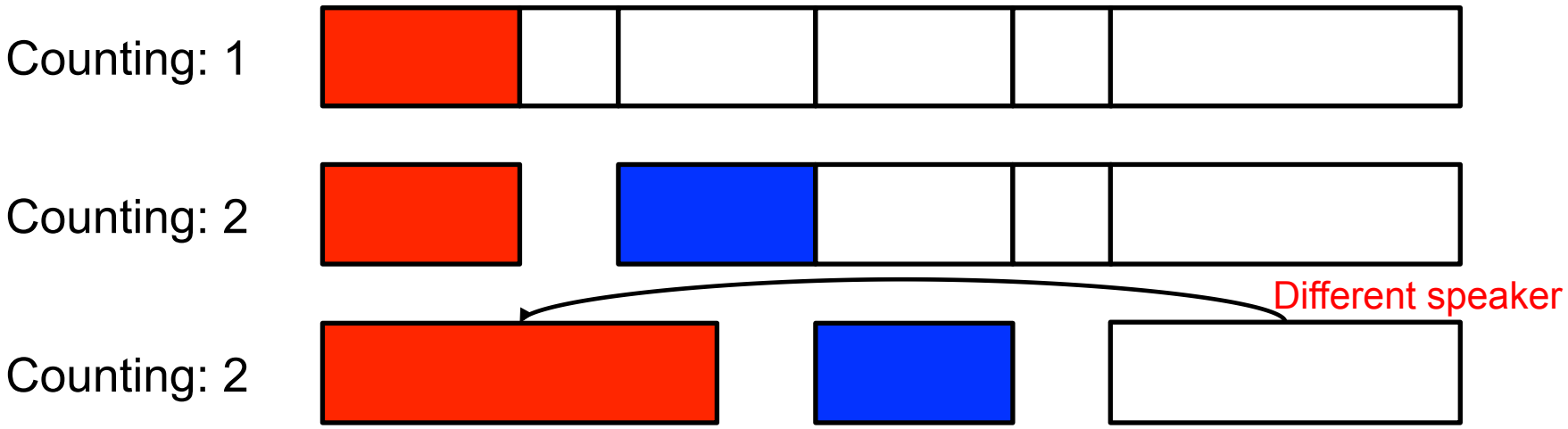
Unsupervised speaker counting



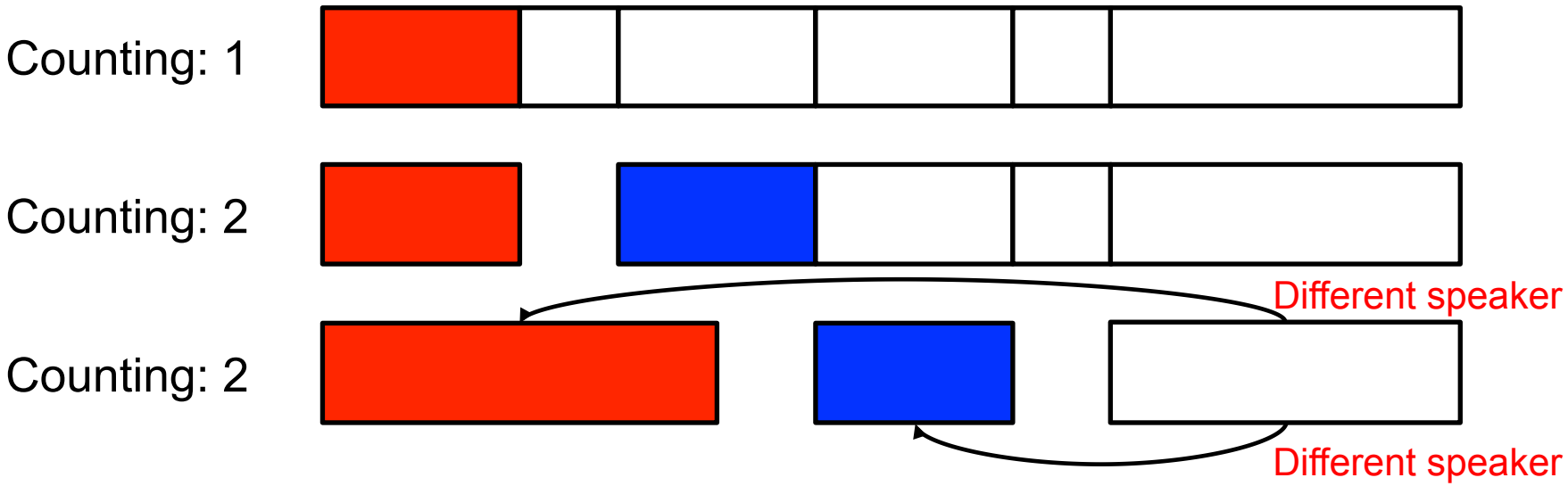
Unsupervised speaker counting



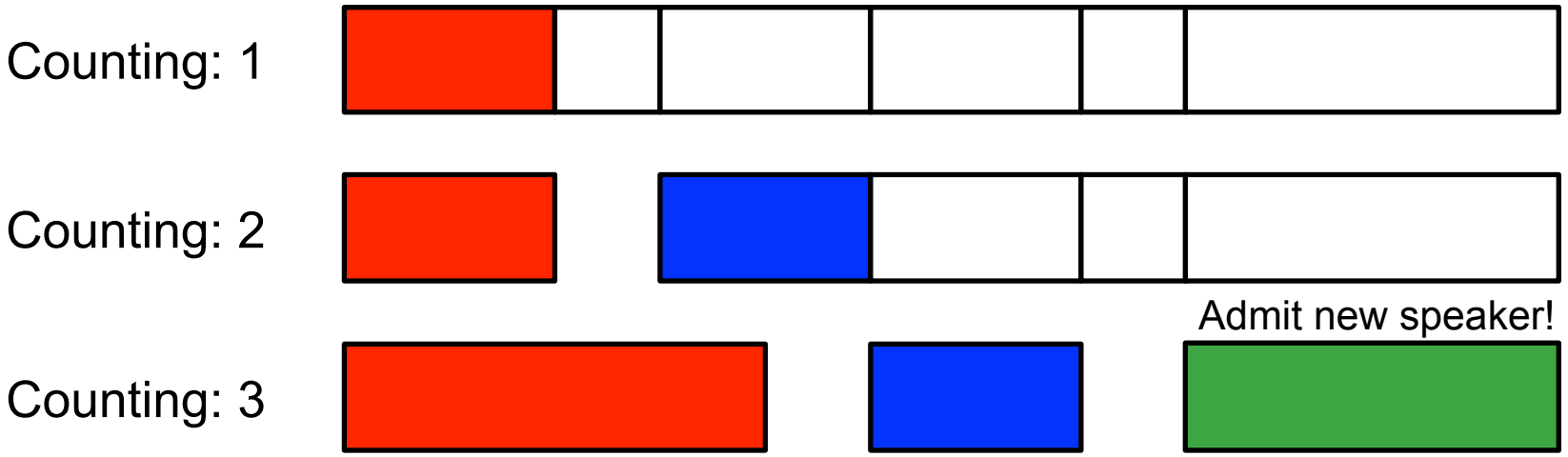
Unsupervised speaker counting



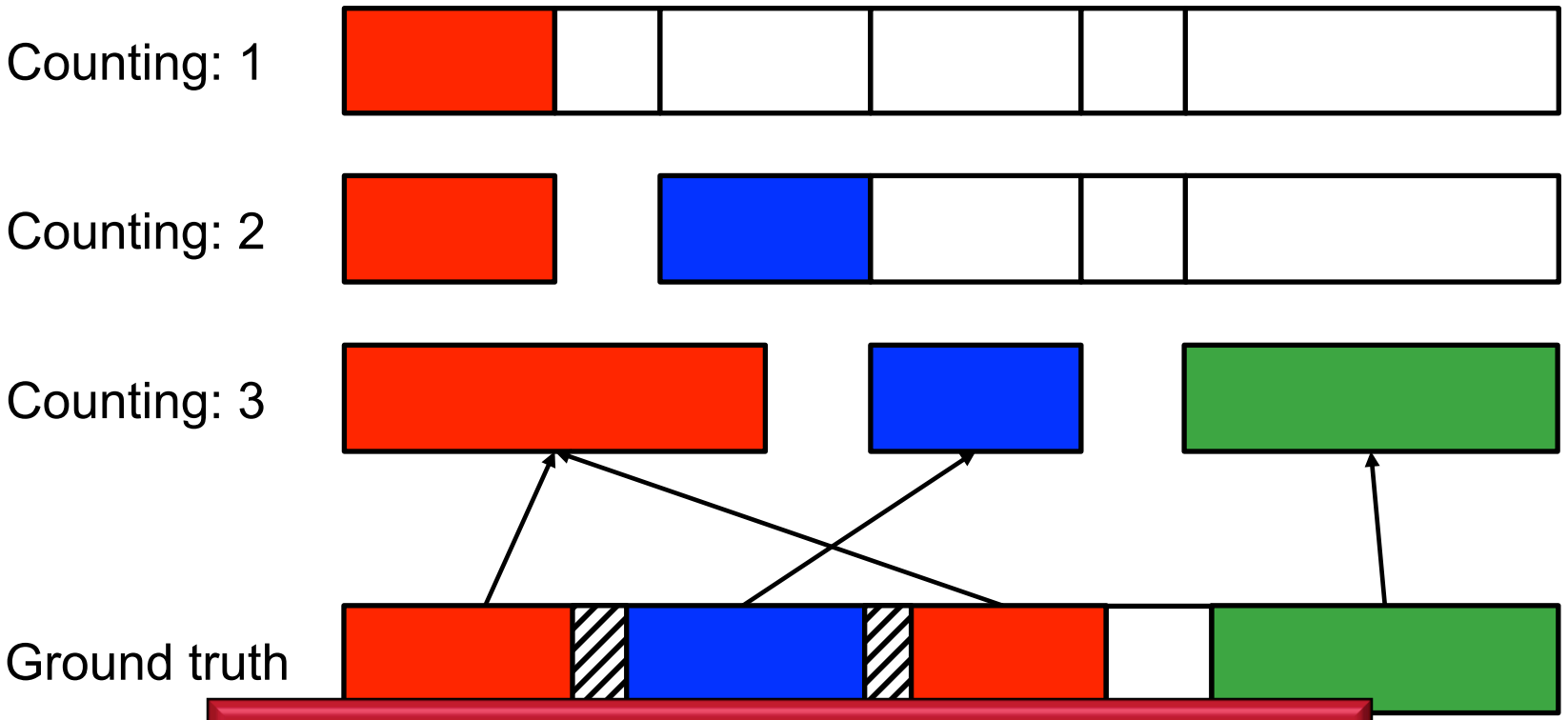
Unsupervised speaker counting



Unsupervised speaker counting



Unsupervised speaker counting

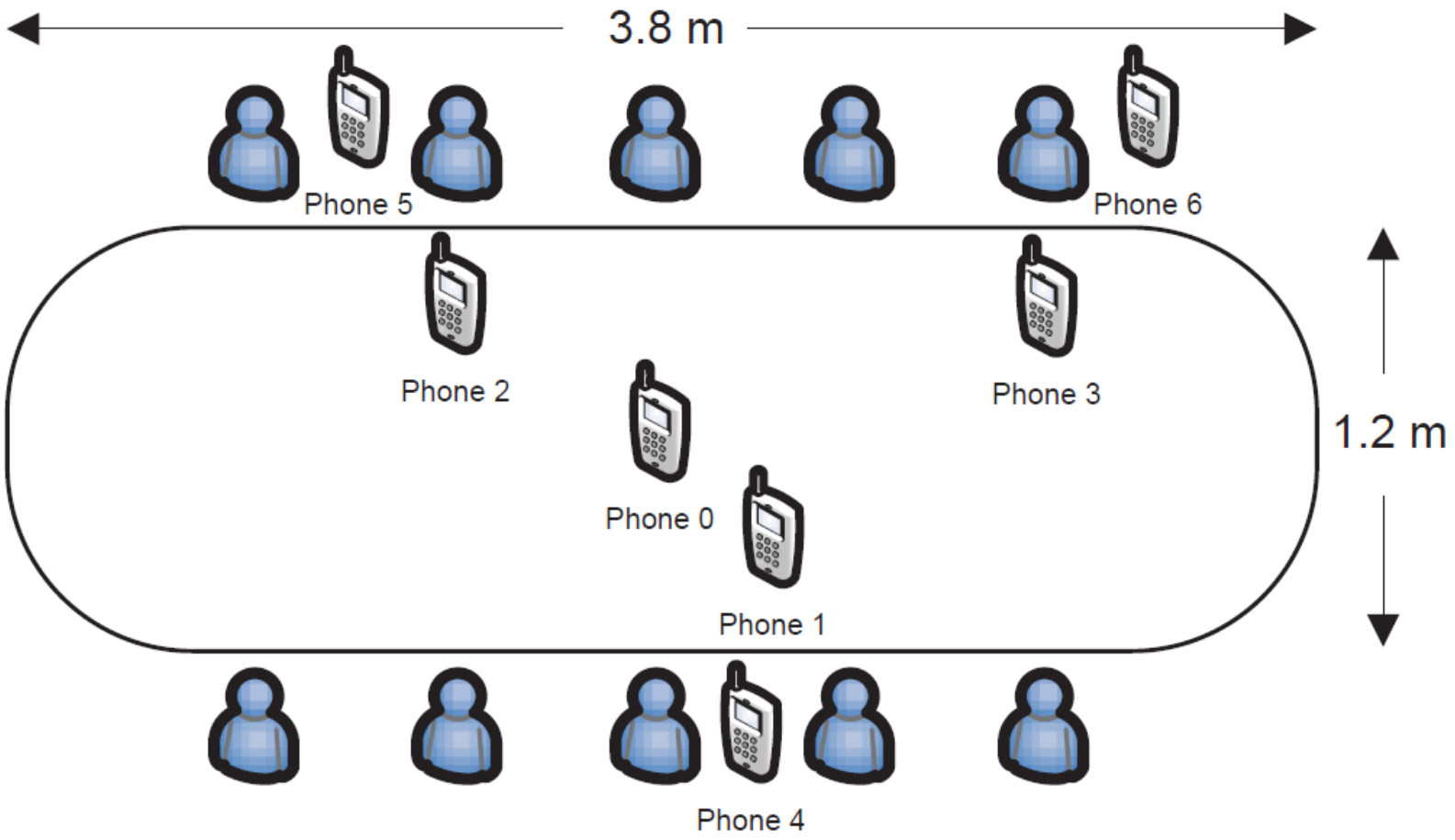


We have three speakers in this conversation!

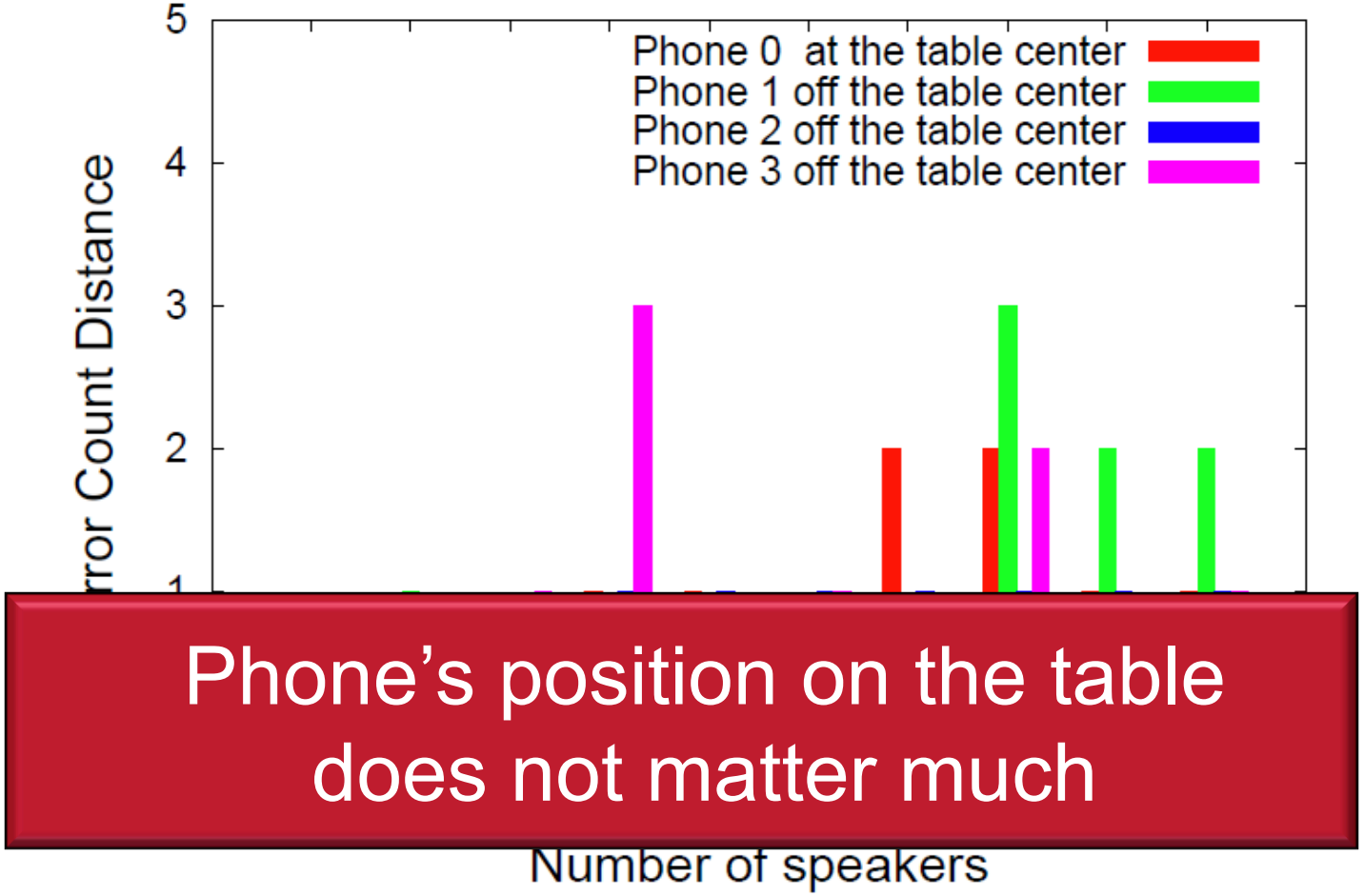
Evaluation metric

- Error count distance
  - The difference between the estimated count and the ground truth.

Benchmark results

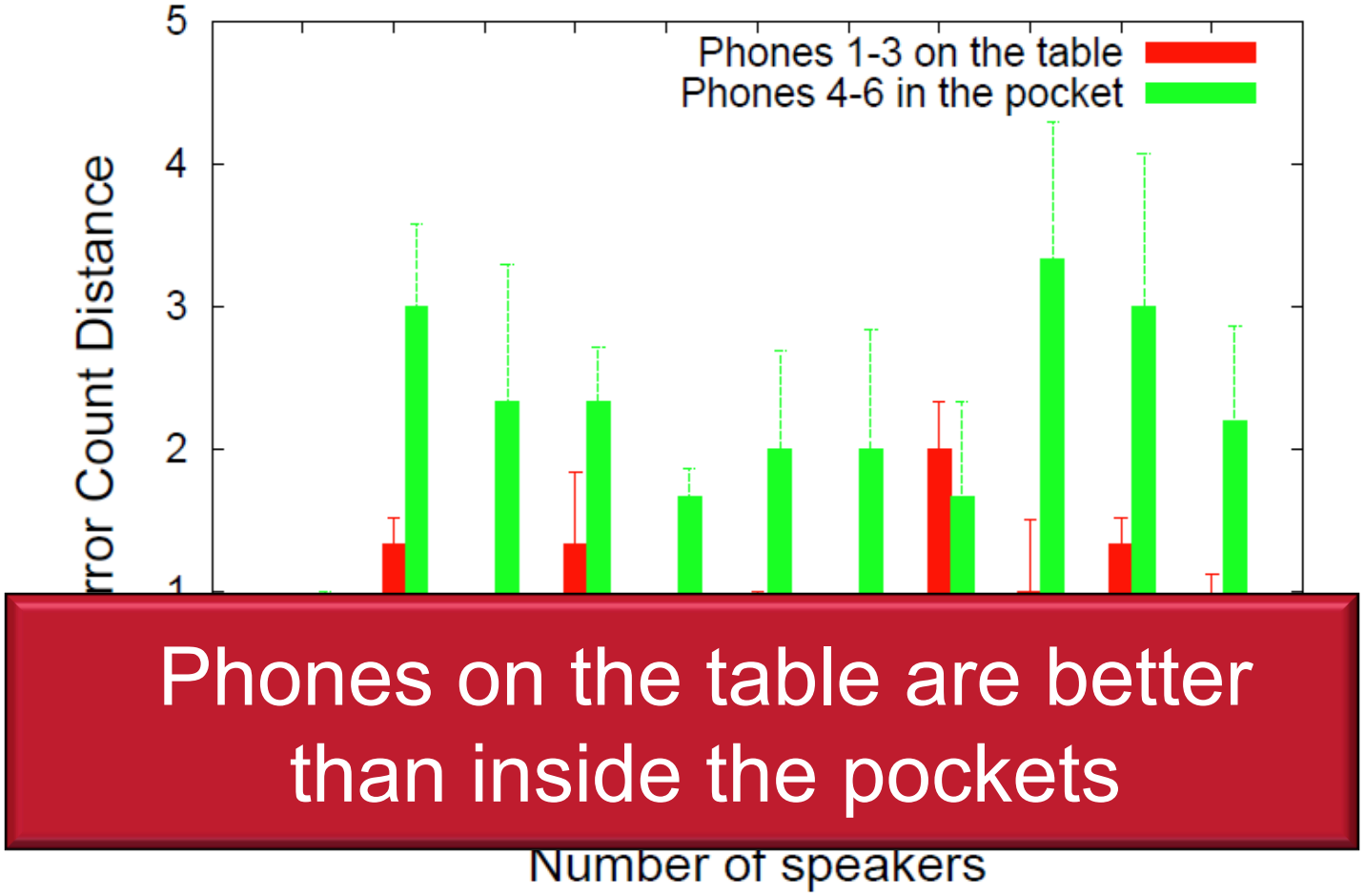


Benchmark results



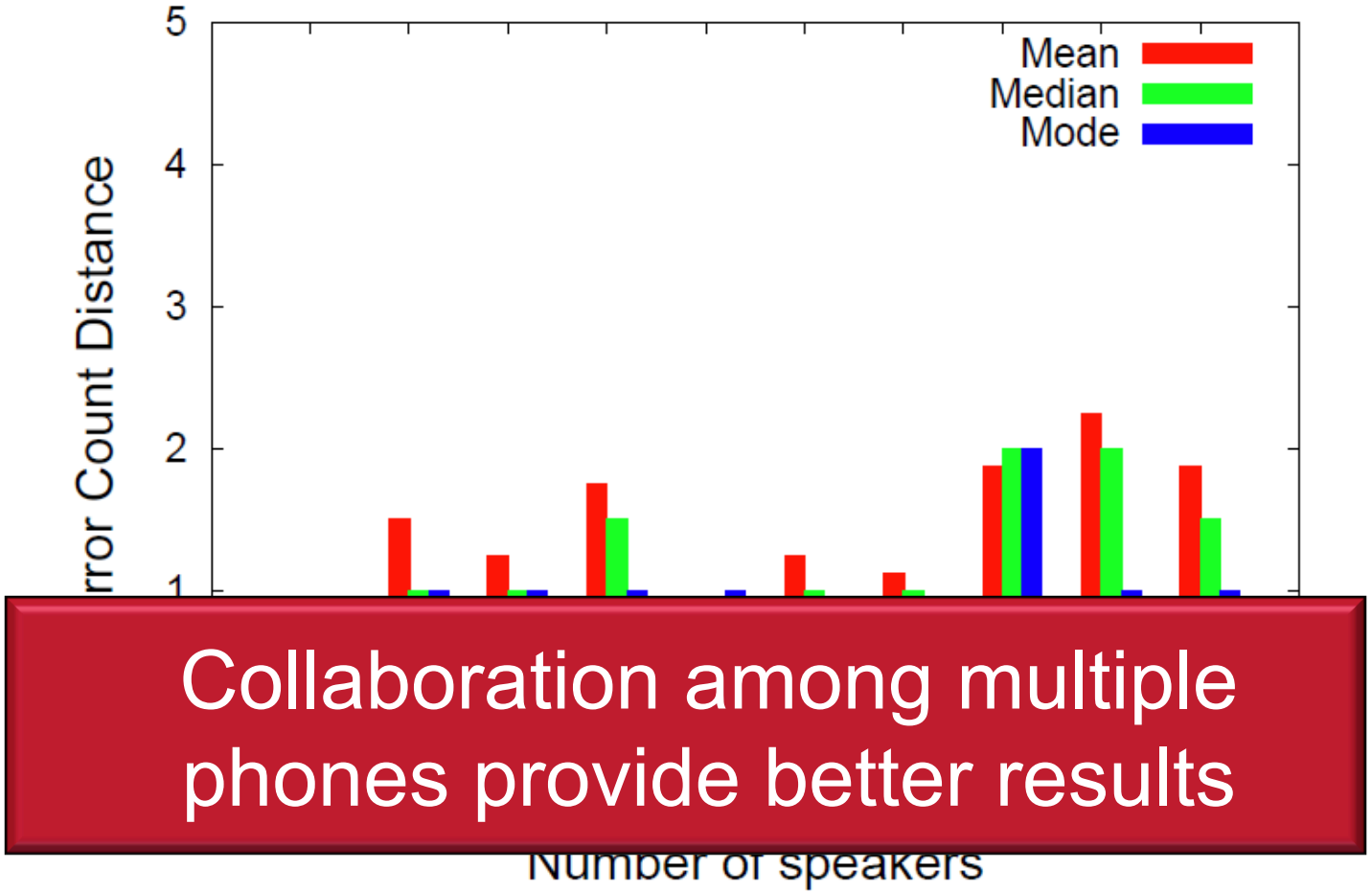
Phone's position on the table does not matter much

Benchmark results



Phones on the table are better than inside the pockets

Benchmark results



Collaboration among multiple phones provide better results

Large scale crowdsourcing effort

- 120 users from university and industry contribute 109 audio clips of 1034 minutes in total.

Private indoor

Public indoor

Outdoor

Large scale crowdsourcing results

	Sample number	Error count distance
Private indoor	**40**	**1.07**
Public indoor	**44**	**1.35**
Outdoor	**25**	**1.83**

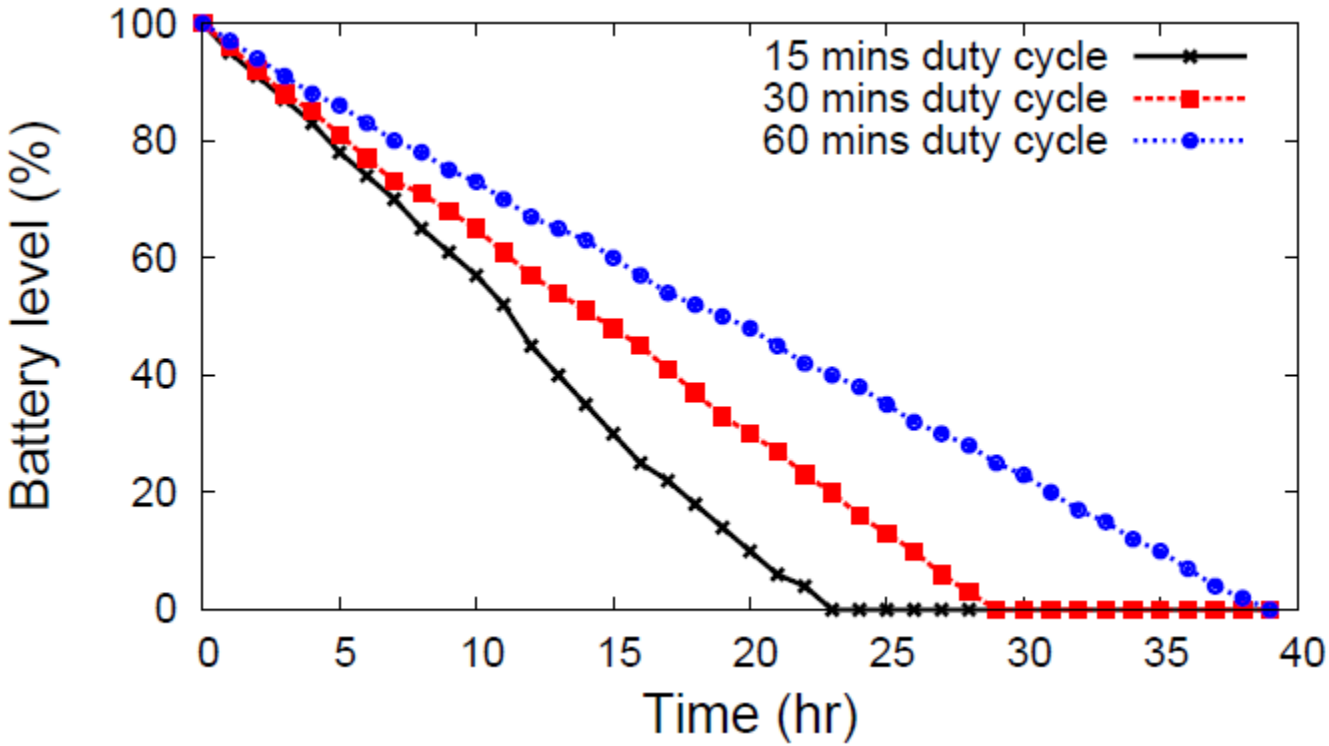
The error count results in all environments are reasonable.

Computational latency

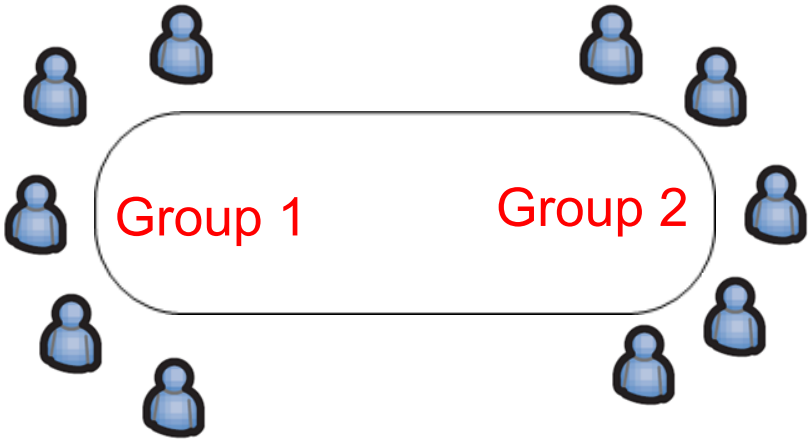
Latency (msec)	HTC EVO 4g	Samsung Galaxy S2	Samsung Galaxy S3	Google Nexus 4	Google Nexus 7
MFCC	42.90	36.71	24.41	22.86	23.14
Pitch	102.71	80.36	58.11	47.93	58.33
Count	175.16	150.47	89.01	83.53	70.23
Tota					

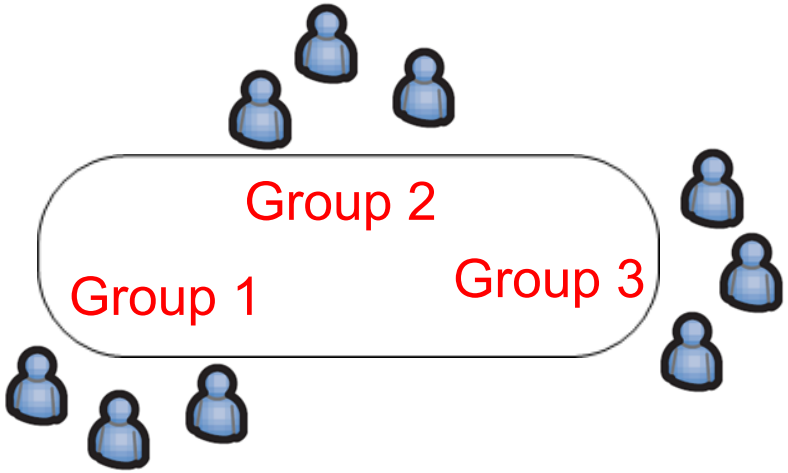
It takes less than 1 minute to process a 5-minute conversation.

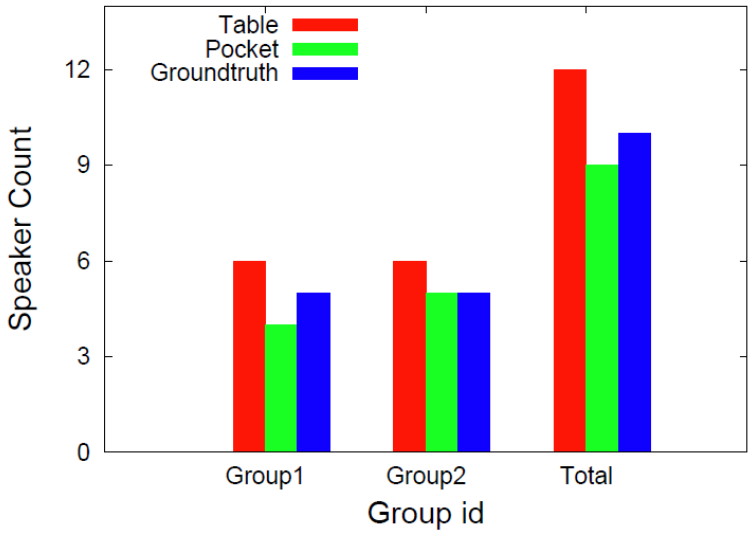
Energy efficiency

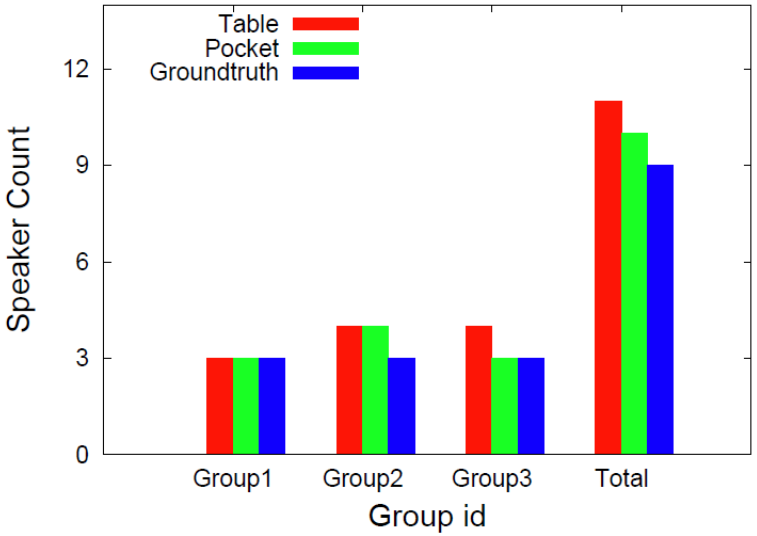


Use case 1: Crowd estimation



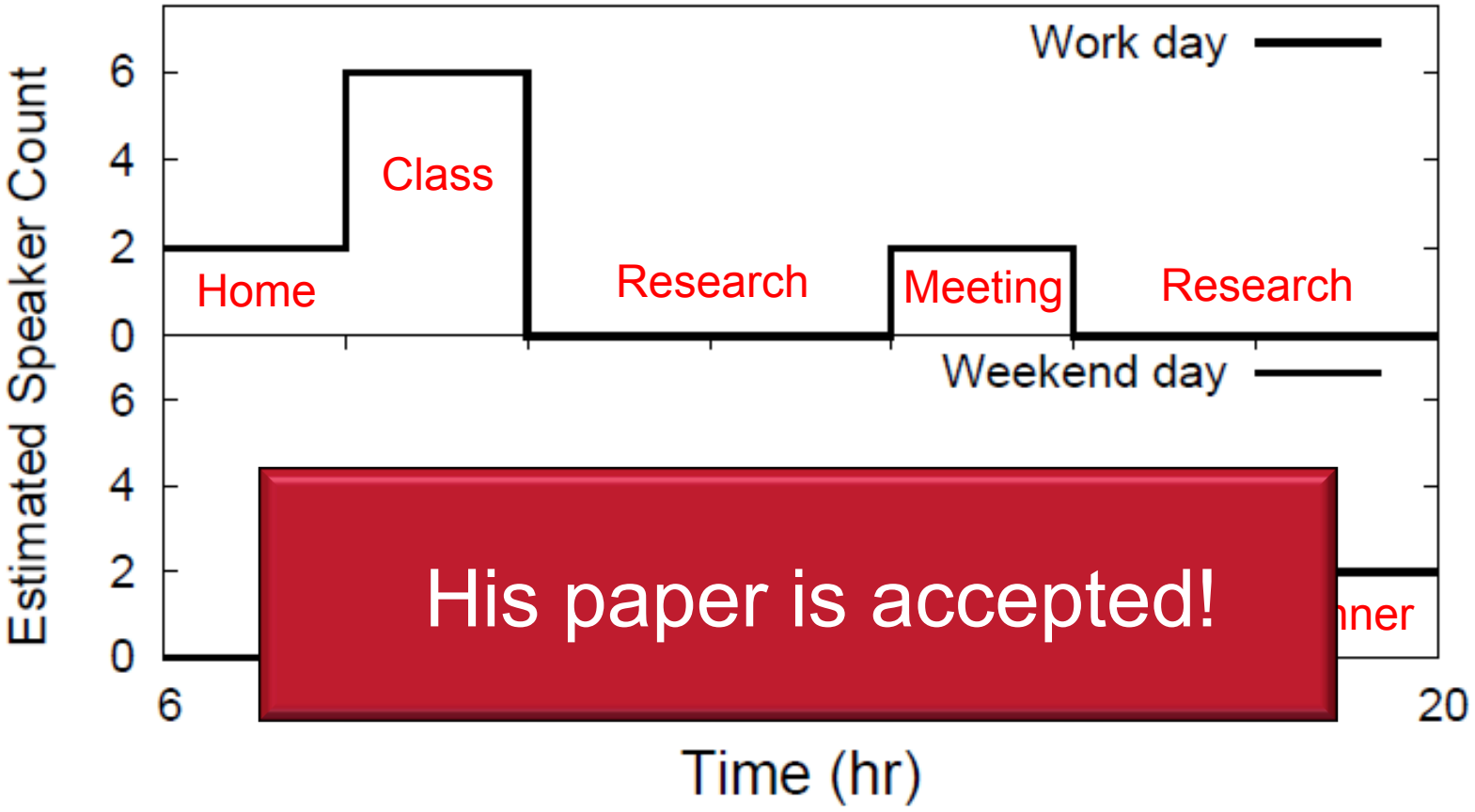




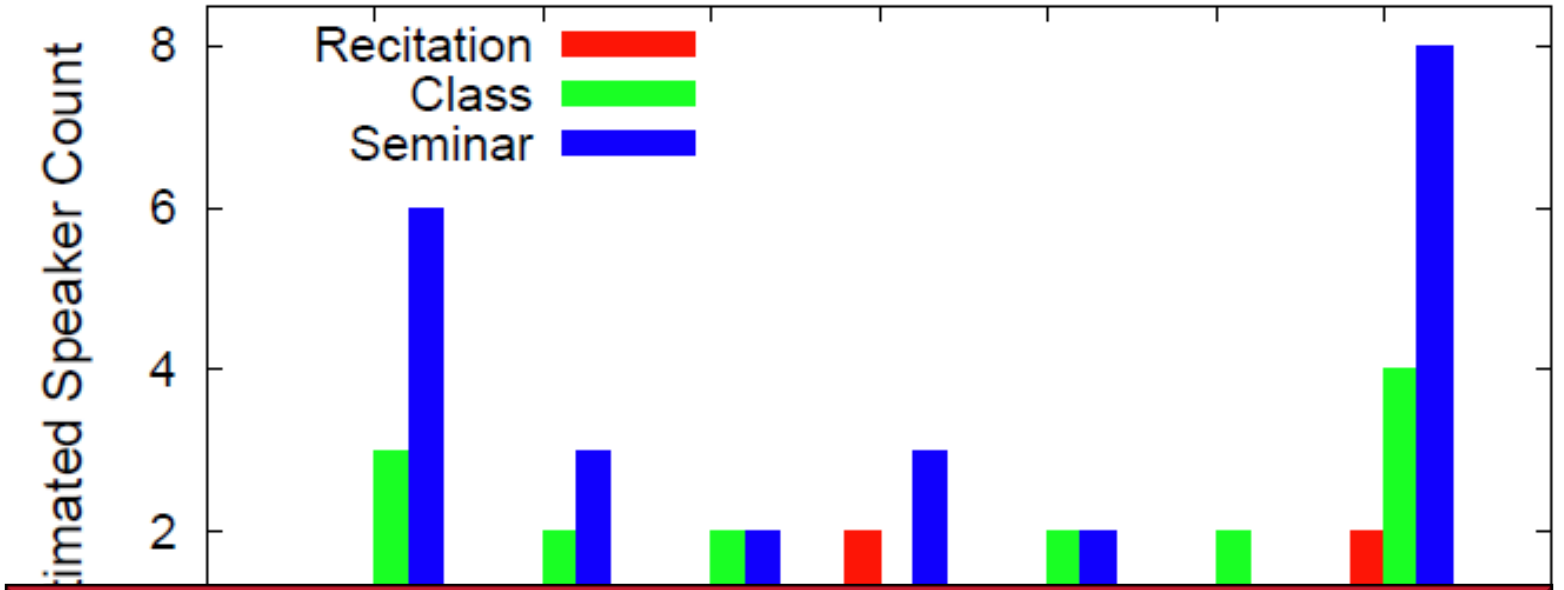


Use case 2: Social log

Ph.D student life snapshot before UbiComp'13 submission



Use case 3: Speaker count patterns



Different talks show very different speaker count patterns as time goes.

Conclusion

- Smartphones can count the number of speakers with reasonable accuracies in different environments.
- Crowd++ can enable different social sensing applications.

Thank you

Chenren Xu
WINLAB/ECE
Rutgers University

Sugang Li
WINLAB/ECE
Rutgers University

Gang Liu
CRSS
UT Dallas

Yanyong Zhang
WINLAB/ECE
Rutgers University

Emiliano Miluzzo
AT&T Labs
Research

Yih-Farn Chen
AT&T Labs
Research

Jun Li
Interdigital
Communication

Bernhard Firner
WINLAB/ECE
Rutgers University