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
Scenario 1: Dinner time, where to go?

I will choose this one!
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

Necklace

Watch

Glass

Brace

Phone
What we already have

Next thing: Voice-centric sensing
Voice-centric sensing

Speech recognition

Family life

Bob

Alice

Speaker identification

Stressful

Emotion detection

Speaker count

2

3
## How to count?

<table>
<thead>
<tr>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>No prior knowledge of speakers</td>
</tr>
<tr>
<td>Background noise</td>
</tr>
<tr>
<td>Speech overlap</td>
</tr>
<tr>
<td>Energy efficiency</td>
</tr>
<tr>
<td>Privacy concern</td>
</tr>
</tbody>
</table>
How to count?

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>No prior knowledge of speakers</td>
<td>Unique features extraction</td>
</tr>
<tr>
<td>Background noise</td>
<td>Frequency-based filter</td>
</tr>
<tr>
<td>Speech overlap</td>
<td>Overlap detection</td>
</tr>
<tr>
<td>Energy efficiency</td>
<td>Coarse-grained modeling</td>
</tr>
<tr>
<td>Privacy concern</td>
<td>On-device computation</td>
</tr>
</tbody>
</table>
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 $\theta$ to capture the distance between speech segments.
Speaker features

- MFCC + cosine similarity distance metric

- Alice’s MFCC in speech segment 3
- Bob’s MFCC in speech segment 2
- Bob’s MFCC in speech segment 1

\( \theta_d > \theta_s \)
Speaker features

- MFCC + cosine similarity distance metric

- Histogram of $\theta_s$
- Histogram of $\theta_d$

1 second speech segment
2-second speech segment
3-second speech segment
10-second speech segment

We use 3-second for basic speech unit.

10 seconds is not natural in conversation!
Speaker features

- MFCC + cosine similarity distance metric

3-second speech segment

![Diagram showing two overlapping bell curves with labels 15 and 30]
Speaker features

- Pitch + gender statistics

![Pitch Distribution Graph](image)
Same speaker or not?

**IF** MFCC cosine similarity score < 15

**AND**

Pitch indicates they are same gender

**ELSEIF** MFCC cosine similarity score > 30

**OR**

Pitch indicates they are different genders

**ELSE**

Not sure
Conversation example

- Speaker A
- Speaker B
- Speaker C

Overlap
Overlap
Silence

Conversation

Time
Unsupervised speaker counting

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

Ground truth

Segmentation

Pre-clustering

Same speaker
Unsupervised speaker counting

- **Ground truth**
- **Segmentation**
- **Pre-clustering**

![Diagram showing the process of unsupervised speaker counting]

- **Ground truth**
  - Red
  - Blue
  - Green
- **Segmentation**
  - Multiple segments
- **Pre-clustering**
  - Multiple clusters
- **Merge**
  - Combining segments and clusters
Unsupervised speaker counting

Ground truth

Segmentation

Pre-clustering

Keep going …
Unsupervised speaker counting

Ground truth

Segmentation

Pre-clustering

Same speaker
Unsupervised speaker counting

Ground truth

Segmentation

Pre-clustering

Merge
Unsupervised speaker counting

Ground truth

Segmentation

Pre-clustering

Keep going …
Unsupervised speaker counting

Ground truth

Segmentation

Pre-clustering

Same speaker
Unsupervised speaker counting

Ground truth

Segmentation

Pre-clustering

Merge
Unsupervised speaker counting

- Ground truth
  - Segmentation
    - Pre-clustering
      - Merge
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

Counting: 1

Not sure
Unsupervised speaker counting

Counting: 1

Drop it

Counting: 1
Unsupervised speaker counting

Counting: 1

Different speaker

Counting: 1
Unsupervised speaker counting

Counting: 1

Counting: 2

Admit new speaker!
Unsupervised speaker counting

<table>
<thead>
<tr>
<th>Counting: 1</th>
<th>Counting: 2</th>
<th>Counting: 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Same speaker
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 2

Same speaker

Different speaker
Unsupervised speaker counting

Counting: 1

Counting: 2

Merge

Counting: 2
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 2

Not sure
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 2
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 2

Drop
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 2

Different speaker
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 2

Different speaker

Different speaker
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 3

Admit new speaker!
Unsupervised speaker counting

Counting: 1

Counting: 2

Counting: 3

Ground truth

We have three speakers in this conversation!
Evaluation metric

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

<table>
<thead>
<tr>
<th>Environment</th>
<th>Sample number</th>
<th>Error count distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private indoor</td>
<td>40</td>
<td>1.07</td>
</tr>
<tr>
<td>Public indoor</td>
<td>44</td>
<td>1.35</td>
</tr>
<tr>
<td>Outdoor</td>
<td>25</td>
<td>1.83</td>
</tr>
</tbody>
</table>

The error count results in all environments are reasonable.
## Computational latency

<table>
<thead>
<tr>
<th>Latency (msec)</th>
<th>HTC EVO 4g</th>
<th>Samsung Galaxy S2</th>
<th>Samsung Galaxy S3</th>
<th>Google Nexus 4</th>
<th>Google Nexus 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>42.90</td>
<td>36.71</td>
<td>24.41</td>
<td>22.86</td>
<td>23.14</td>
</tr>
<tr>
<td>Pitch</td>
<td>102.71</td>
<td>80.36</td>
<td>58.11</td>
<td>47.93</td>
<td>58.33</td>
</tr>
<tr>
<td>Count</td>
<td>175.16</td>
<td>150.47</td>
<td>89.01</td>
<td>83.53</td>
<td>70.23</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It takes less than 1 minute to process a 5-minute conversation.
Energy efficiency
Use case 1: Crowd estimation

Group 1   Group 2

Group 2   Group 1   Group 3

Speaker Count

Group id

Table  Pocket  Groundtruth

Group 1  Group 2  Total

Speaker Count

Group id

Table  Pocket  Groundtruth

Group 1  Group 2  Group 3  Total
Use case 2: Social log

Ph.D student life snapshot before UbiComp’13 submission

Estimated Speaker Count

Time (hr)

His paper is accepted!
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