Crowd++: Unsupervised Speaker Count on Smartphones

Chenren Xu
WINLAB and Dept. of ECE, Rutgers University
iab@winlab.rutgers.edu
May 15th, 2014
Scene 1: Dinner time, where to go?
Scene 2: Is your kid social?
Scene 3: Which class is engaging?
Speaker Count

- Dinner time, where to go?
  - Find the place where most people are 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
How to count?

Challenges

No prior knowledge of speakers
Background noise
Energy efficiency
Privacy concern

Solutions

Speaker dependent features extraction
Frequency-domain filter
Coarse-grained modeling
On-device computation
Crowd++ overview

Step 1: Microphone recording.

Step 2: Make the raw audio signal stationary.

Step 3: Filter the unvoiced data.

Step 4: Extract speaker dependent features.

Step 5. Count the speakers using all the above information.

Source code: https://github.com/lendlice/crowdpp
Speech detection

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

![Diagram showing the frequency range of human voice from 50 Hz to 450 Hz]
Speaker features

- MFCC (Mel-frequency cepstral coefficients)
  - Speaker identification/verification
    - Alice or Bob, or else?
  - Speaker counting
    - No prior information is available
Speaker features

- MFCC + cosine similarity distance metric

We use the angle $\theta$ to capture the distance between speech segments.
Speaker features

- MFCC + cosine similarity distance metric

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

10-second utterance is not common in conversation!
Speaker features

- MFCC + cosine similarity distance metric

3-second speech segment

Thresholds trade-off the sensitivity to admitting new speaker, as well as filtering overlap/silence.
Speaker features

- Pitch + gender statistics

![Pitch vs Gender Diagram](image)

- He is male!
- She is female!
Same speaker or not?

\[
\text{IF } \text{MFCC cosine similarity score } < \theta_s \\
\quad \text{AND} \\
\text{Pitch indicates they are same gender}
\]

\[
\text{ELSEIF } \text{MFCC cosine similarity score } > \theta_u \\
\quad \text{OR} \\
\text{Pitch indicates they are different genders}
\]

Only admit new speaker when its speech is different from all the admitted speakers.
Conversation example

Speaker A

Speaker B

Speaker C

Conversation

Time

Overlap

Overlap

Silence
Unsupervised speaker counting

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

Ground truth

Segmentation

Pre-clustering

Pre-clustering

Pre-clustering

Pre-clustering

Pre-clustering

Merge speaker

Keep merging
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.
We have three speakers in this conversations!
Single group benchmarking results

10 participants talk one by one in sequence.

7 smartphones are placed on the table and in the pockets.

Phone’s position does not matter much.

It’s better to place the phone on the table.

Collaboration provides better results.
Multiple groups benchmarking results

Group 1  Group 2

Group 2  Group 1  Group 3

Speaker Count

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Table</th>
<th>Pocket</th>
<th>Groundtruth</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Speaker Count

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Table</th>
<th>Pocket</th>
<th>Groundtruth</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Crowdsourcing effort

- 120 users contribute 109 audio clips of 1034 minutes

<table>
<thead>
<tr>
<th>Private Indoor Environments</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker #</td>
<td>Sample #</td>
<td>Place</td>
<td>AECD</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>Home</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>Office</td>
<td>0.82</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>Office</td>
<td>1.25</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>Kitchen</td>
<td>1.28</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>Kitchen</td>
<td>2</td>
</tr>
<tr>
<td>Overall</td>
<td>40</td>
<td>Quiet Indoor</td>
<td>1.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Public Indoor Environments</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker #</td>
<td>Sample #</td>
<td>Place</td>
<td>AECD</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Restaurant</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>Food Court</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>Coffee Shop</td>
<td>1.86</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>Shopping Mall</td>
<td>1.82</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>Super Market</td>
<td>1.58</td>
</tr>
<tr>
<td>Overall</td>
<td>44</td>
<td>Noisy Indoor</td>
<td>1.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outdoor Environments</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker #</td>
<td>Sample #</td>
<td>Place</td>
<td>AECD</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>Plaza</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>Parking Lot</td>
<td>1.2</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>Plaza</td>
<td>2.29</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Parking Lot</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>Patio</td>
<td>2.67</td>
</tr>
<tr>
<td>Overall</td>
<td>25</td>
<td>Noisy Outdoor</td>
<td>1.83</td>
</tr>
</tbody>
</table>

Average error count distance:

- 1.07
- 1.35
- 1.83
Lessons learned

- Speaker count can be estimated on the smartphone in an unsupervised manner.
- Speaker counting technique can enable a number of social sensing applications.
- Accuracies: private indoor > public indoor > outdoor
- We need low-cost noise cancellation technique to improve the accuracy.