Speaker Recognition: Building the Mixer 4 and 5 Corpora

Linda Brandschain, Christopher Cieri, David Graff, Abby Neely, Kevin Walker
{brndschn|ccieri|graff|aneely|walkerk}@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium
Motivation

◆ Mixer supports R&D of speaker recognition systems robust to variation in:
  ● language: Arabic, Mandarin, Russian, Spanish
  ● channel: telephone + 8 to 14 microphones
  ● conversational situation: telephone conversation, interviews, reading words, phrases, sentences, transcripts, written texts

◆ Mixer 4
  ● channel variation

◆ Mixer 5
  ● channel
  ● conversational situation
## Comparison of Phases

<table>
<thead>
<tr>
<th></th>
<th>SB</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core Calls (8+)</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Variable Environments</strong></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unique Handset (4+)</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Extended Data (20+)</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Multilingual (4+)</strong></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cross Channel (2 or 4)</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Transcript Reading (2+)</strong></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Interviews (6)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Mixer Platform Design

- Mixer platform designed to address changing telephony

  - Issues Encountered
    - increased cell phone use
    - inexpensive domestic and international calling rates
    - rise in use of call forwarding and call-screening

  - Solutions
    - reduce hours of the study
    - exploit all lines available to robot operator
    - reduce impediments to matching subjects
      - allow any pairing, including duplicates
    - over recruit
    - set goals 20 – 25% higher than required by project sponsors
    - lower per call payment; large completion bonuses
    - encourage subjects to give true, narrow availability schedule
    - increase robot activity to combat increased miss ratio
Protocol
Protocol
Mixer Call Platform

- Mixer 4 & 5 conducted simultaneously
- Studies began when participant pool >= 200
- 40 topics cycled
  - current political and social issues, religion, hobbies, sports, etc
  - no penalty for speaking “off topic” so long as conversation is topical
  - participants could refuse call after hearing the topic of the day

- Auditing
  - calls audited for length, sound quality, quantity/suitability of speech.
  - participants who reached their goal were deactivated
## Multi-Channel Set-Up

<table>
<thead>
<tr>
<th>Ch</th>
<th>Microphone</th>
<th>Placement</th>
<th>Subject/Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shure MX185 Lavalier</td>
<td>Interviewer</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Shure MX185 Lavalier</td>
<td>Subject</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Etymotic Micro-array</td>
<td>Interviewer</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Shure MX418X Podium</td>
<td>Desk Front</td>
<td>Center</td>
</tr>
<tr>
<td>5</td>
<td>Crown PZM-6D</td>
<td>Desk Top</td>
<td>Center</td>
</tr>
<tr>
<td>6</td>
<td>Audio Technica AT3035</td>
<td>Desk Front</td>
<td>Right</td>
</tr>
<tr>
<td>7</td>
<td>Audio Technica Pro45</td>
<td>Hanging</td>
<td>Center</td>
</tr>
<tr>
<td>8</td>
<td>Panasonic Camcorder</td>
<td>Desk Top</td>
<td>Right</td>
</tr>
<tr>
<td>9</td>
<td>RODE NT6</td>
<td>Desk Front</td>
<td>Far Left</td>
</tr>
<tr>
<td>10</td>
<td>RODE NT6</td>
<td>Desk Front</td>
<td>Center Left</td>
</tr>
<tr>
<td>11</td>
<td>RODE NT6</td>
<td>Desk Front</td>
<td>Center Right</td>
</tr>
<tr>
<td>12</td>
<td>RODE NT6</td>
<td>Desk Front</td>
<td>Center Far Right</td>
</tr>
<tr>
<td>13</td>
<td>AcoustiMagic Array</td>
<td>Wall Mounted</td>
<td>Center</td>
</tr>
<tr>
<td>14</td>
<td>Lightspeed Headset</td>
<td>Subject</td>
<td></td>
</tr>
</tbody>
</table>
Mixer 4

- Mixer 4 was designed to support speaker recognition research and technology evaluations

Demographics of Subject Pool
- Native Speakers of American English
- 25% from Philadelphia
- 25% from Berkeley
- 50% from the entire US, however we recruited heavily in Georgia, Texas, Illinois, and New York

Original Goals for Mixer 4
- 400 Subjects that made 10, 10 minute phone calls
- 200 Visited one of our two sites where they completed 2 cross-channel call
- 100 Participants were asked to complete extended data calls (20 x 10-minute phone calls)
Mixer 4 Call Yields

- Total Calls: 233
- Total Minutes: 17,200
- Total Hours: 287
- Subjects with 10+ Calls: 233
- Subjects with 20+ Calls: 52

Calls Made vs. Speakers

- X-axis: Calls Made
- Y-axis: Speakers

LREC 2008, May 26 – June 1, Marrakesh
Mixer 5

- Mixer 5 focused on cross-channel recordings of face to face interviews where the goal is to elicit speech within a variety of situations.

- Demographics of Subject Pool
  - Native language undefined, however participants had to be fluent in English
  - Approximately 50% recruited from Philadelphia, PA
  - Approximately 50% recruited from Berkeley, CA

- Goals for Mixer 5
  - 300 Participants
  - Each Participant must complete 6 half hour sessions completed in no less than 6 days. Each session had a mandatory 30 minute break between sessions.
  - Each of the 300 Participants must also complete 10 ten-minute phone calls
  - Foreign language calls were encouraged but not required
  - Bonuses were issued for the completion of 4 unique phone calls

- High/Low Vocal Effort Phone Calls
  - ~1/3 of Mixer 5 Participants completed these calls
  - Lightspeed XLC-20 headphones provide 40db passive acoustic isolation
  - High Vocal Effort: Input audio is 65dB and relative levels of the mix components are 30% side-tone, 40% remote speaker and 30% white noise.
  - Low Vocal Effort: Input audio is 65dB with no white noise.
<table>
<thead>
<tr>
<th>Session Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeating Questions</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Warm-up</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Family Personal</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Informal Conversation</td>
<td>20</td>
<td>9</td>
<td>14</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>70</td>
</tr>
<tr>
<td>Transcript Reading</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>Story Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Sentence Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Phrase/Word List Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Low Vocal/Effort</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>High Vocal/Effort</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Total Session</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>180</td>
</tr>
</tbody>
</table>
Participant Screen ON. The participant reads these phrases and words in a normal reading voice.

March thirtieth
Mixer 5 Call Yields

Total Calls: 2919
Total Minutes: 14595
Total Hours: 243
Subjects with 10+ Calls: 245
Future Work

◆ Mixer 1 & 2
  ● in LDC publication pipeline
◆ Mixer 3
  ● used in SRE06 & LRE07; remainder reserved for future evaluation
◆ Mixer 4
  ● collection underway
  ● part used in SRE08 remainder reserved for future evaluation
◆ Mixer 5
  ● interview collection ahead of schedule
  ● phone call collection also well underway
  ● part used in SRE08; remainder reserved for future evaluation
◆ Mixer 6 (Graybeard)
  ● subjects from previous CTS collection return to join
◆ Potential new studies
  ● conduct Mixer 5 style interviews in other languages
  ● conduct studies like Mixer 1 & 2 but involving other languages
◆ All Mixer data will be published after its use in technology evaluations.