Dimensions of Speaker Recognition Research Data
Christopher Cieri, Mark Liberman
University of Pennsylvania, Linguistic Data Consortium
NIST Speaker Recognition Systems

- systematic exploration of technology challenges
- i.e. text, channel, room, language independence
- supporting data consists of multiple samples per talker

- varying and controlling for variation in:
  - talkers
  - sessions
  - communicative situation (style)
  - environment and including interlocutor
  - sensors
  - transmission channels
  - and of course linguistic variety
LDC Roles

- distribution & archiving (CD ➔ DVD ➔ HD ➔ Cloud ➔ Grid)
- language resource production, including quality control
- intellectual property rights and license management
- human subject protocol management
- data collection
- annotation and lexicon building
- creation of tools, specifications, best practices
- knowledge transfer: documentation, metadata, consulting, training
- corpus creation research (meta-research) and academic publication
- resource coordination in large multisite programs
- workshop organization
- service to multiple research communities
  - funding panelists, workshop participants, oversight committee members
- funder (grants in data program): 4 years, 70 corpora, 41 recipients, $128,000
<table>
<thead>
<tr>
<th></th>
<th>96</th>
<th>97</th>
<th>98</th>
<th>99</th>
<th>00</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRE</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SRE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BN Re</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CTS Re</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SDR</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TDT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ACE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DUC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>STD</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MetricsMaTr</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HaRT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TAC KBP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TREC Vid SED</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TREC Vid MED</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TREC Vid MER</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
SR Corpus Building

- Planning among developers, sponsors, evaluation and data teams
- Recruitment
  - demographics targeted to research needs
  - note availability
- Collection
  - Calls
    - robot operator calls subjects at their available times, subjects can call toll-free
    - different topics suggested each day
    - rules for pairing talkers vary by study
  - Interviews
    - vary activities, rooms, sensors
- Annotation
  - speaker ID, sound quality, topic, interview segments
- Monitoring: monitor progress and adjust practice
- Publication: final LDC QC, NIST QC & sampling for test data,

Talkers

- universal contributor database, unique ID, no SPII shared
- new or repeating
- demographic selection, not just metadata
  - sex, age, region (dialect), ethnicity
  - monolingual and multilingual, speaking in other or multiple languages
- intrinsic variation
  - aging
  - communicative situation
  - language spoken
- contacted via: social network, community, senior and immigrant centers, Craig’s list, email, email lists, web, handbill, poster, newspaper, radio and, MTurk
- incentivized: money, socializing, ‘therapy’, etc.
 Sessions

- date/time: controlled, scheduled or free
- location: unknown, known
- number: 4, 8, 20, 30
- unique talker combinations
- mediated by
  - phone line, other communication channel, air, no glass
- durations: 5, 6, 10, 20, 30, 60 minutes, unique, not copied
- intersession intervals, sessions per unit time
- session initiated by talker, robot, interviewer
- communicative situation
Communicative Situation

- natural or experimentally manipulated
- conversation, interview, repeating questions, reading words, (shibboleths), digit strings, phrases, (phonetically rich) sentences, transcripts, stories, names (own), twenty questions, map task, Lombard speech
- noise
  - real (affects talker as well) or additive
  - acoustic, electromagnetic, e.g. HVAC, fluorescent light, city-noise
  - hi-/lo- noise eliciting different vocal effort, **but no screaming**
- topic: assigned, free
- distance to interlocutor
- sensor/channel (affects recording but also talker)
- language: (non-)English, monolingual, bilingual
  - ‘Arabic’, Dari, Farsi, Levantine, Mandarin, Pashto, Russian, Spanish, Urdu
Environment & Interlocutor

● real or simulated (afterwards using room modeling software)

● indoors, outdoors, moving vehicle, noisy public space
  ● number of rooms (1-7)
  ● room size, shape, reverberation
    ■ provide impulse response, measurements, photos
      ● clicks, tone sweeps, colored-noise
      ● issues with room comparison/rating
    ■ regularly (daily) ‘calibration’
  ● multiple talker locations within room

● interlocutors
  ● relationship: intimates, familiars, famous (SCOTUS), strangers
  ● naïve or claque (confederate)
  ● human or machine (SPINE)
Audio Recording: Sensors

- Microphones
  - head-mounted, throat, ear bud, ear boom, lavalier, studio, studio instrument, podium, dictaphone, computer, conference room, reference, camcorder, shotgun, array, pilot-headset, pzm, array hearing aid, ‘exotic’

- Handsets
  - wireline, wireless, cell, speaker phone

- unique, repeatable, repeated x times
- pick up pattern, sensitivity, frequency response
- placement: distance, orientation, visible or not
- within operating parameters or not
Audio Recording: Transmission Channels

- captured live or re-transmitted
- number (cross-channel, TSID)
- types
  - telephone
    - POTS (national networks), cell: GSM, TDMA, CDMA
    - typically 4-wire
  - broadband, internet (voip), public radio, walkie talkie, audio chat
  - military channels (SPINE)
- time-alignment
  - via hardware, timecode, worldclock
  - via cross correlation
Metadata & Annotation

- **Metadata**
  - self-reported, judged, deduced
  - personal: height, weight, oral appliance, impairment, language: proficiency
  - session: intelligibility, emotion, deception, noise/vocal effort

- **Audit & Annotation**
  - Speaker ID: confirm pairs of segments from same speaker
    - Need gold standard; need not replicate system decision (HASR)
    - Use name recording, visual ID, content, previous recordings, personal knowledge
    - False alarms rare, misses cannot be easily resolved
  - Topic
  - Transcription
    - human or machine generated
  - Session vs. Segment level: audit decisions only valid for segments judged
<table>
<thead>
<tr>
<th></th>
<th>SB</th>
<th>SB2</th>
<th>SB2</th>
<th>SB2</th>
<th>SB</th>
<th>SB</th>
<th>M1 &amp; 2</th>
<th>M3</th>
<th>M4 &amp; 5</th>
<th>GB</th>
<th>M6</th>
<th>M7</th>
<th>SRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talkers</td>
<td>543</td>
<td>657</td>
<td>679</td>
<td>640</td>
<td>254</td>
<td>419</td>
<td>4800</td>
<td>4050</td>
<td>1452</td>
<td>171</td>
<td>595</td>
<td>434</td>
<td>358</td>
</tr>
<tr>
<td>Sides</td>
<td>5K</td>
<td>7K</td>
<td>9K</td>
<td>5K</td>
<td>3K</td>
<td>4K</td>
<td>28K</td>
<td>20K</td>
<td>6K</td>
<td>2K</td>
<td>9K</td>
<td>11K</td>
<td>4K</td>
</tr>
<tr>
<td>8+ Calls</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>20+ Calls</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Settings</td>
<td>IOV</td>
<td>IOV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Handsets</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Languages</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cell Nets</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channels</td>
<td>8</td>
<td>14</td>
<td>18</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocal Effort</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Longitud.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Other Corpora

- **YOHO (1994):** 138 speakers, *14 sessions*, digit strings
- **King (1995):** 50 male speakers, 2 settings, 2 channels, *task speech*
- **LLHDB (1998):** 53 speakers, *10 handsets*, read & task speech
- **AHUMADA (1998):** 104 speakers, 6 sessions, 16 channels, read & spontaneous speech in Spanish
- **TSID (1999):** ? speakers, 3 sessions, *18 channels*, read & task speech
- **SUSAS (1999):** 32 speakers, *stress conditions*
- **SPINE (2000):** 40 speakers, 420 sessions, *4 noise/channel pairs*, collaborative speech
- **CSLU Sp.Rec. (2006):** 91 speakers, 12 sessions over 2 years, QA & conversation
- **SCOTUS (2008):** oral arguments, known & unknown speakers, changing conditions
- **TM (2011):** 100 speakers, 2 channels including *throat mic*, read speech, non-native
- **VoCMex (2012):** 33 speaker, 3 sessions, 2 channels, *Spanish* read speech
- **RSR2015 (2012):** 298 speakers, 9 sessions, 6 channels, read and task speech
  - pass-phrases, command and control, digit strings
Other Directions

◆ Phanotics
  • quantifying linguistic variation as correlated with idiolect and dialect
  • 297 Fisher/Mixer calls transcribed
  • from subjects self identified as African- and European-American
  • annotated for sociolinguistic variables
  • features used in speaker and dialect ID systems

◆ HASR
  • humans attempting to do speaker recognition as in the NIST evaluations
  • open to all: experts and novices, very few experts contributed
  • using difficult cross-channel trials from Mixer 6 (SRE10)
  • 2 phases, 150 trials total, 20 systems
  • Miss: 35-39%, FA: 41-47%
  • HASR systems did not compare favorably to automatic systems on these trials