SRE10

NIST Speaker Recognition Workshop

Linda Brandschain, Chris Caruso, Christopher Cieri,

David Graff, Abby Neely, Kevin Walker

{brndschn|carusocr|ccieri|graff|neely|walkerk}@ldc.upenn.edu

University of Pennsylvania

Linguistic Data Consortium

Linguistic Data Consortium

Introduction

- **Mixer-6 collection**

  general stats, protocols for recruiting, call-collection, interviews, details on vocal-effort phone calls, call-audit process, overview of audit results, interview audit processes, overview of results data delivery to NIST

- **Greybeard Collection**

  general stats, legacy collections protocols for recruiting, call collection, call and legacy audit processes, overview of results data deliveries to NIST (accounting for majority of "empty segments")

- **SRE08 data (from mixer-3/4/5 collections)**

  general stats per collection, including date ranges, protocol differences relative to GB and Mixer-6, (esp. remote subjects)

- **Summary of problems:**

  "one subj_id => multiple voices", "multiple subj_ids => one voice", bad hard drives, empty speech segments in test set.

Linguistic Data Consortium

Mixer 6 Collection

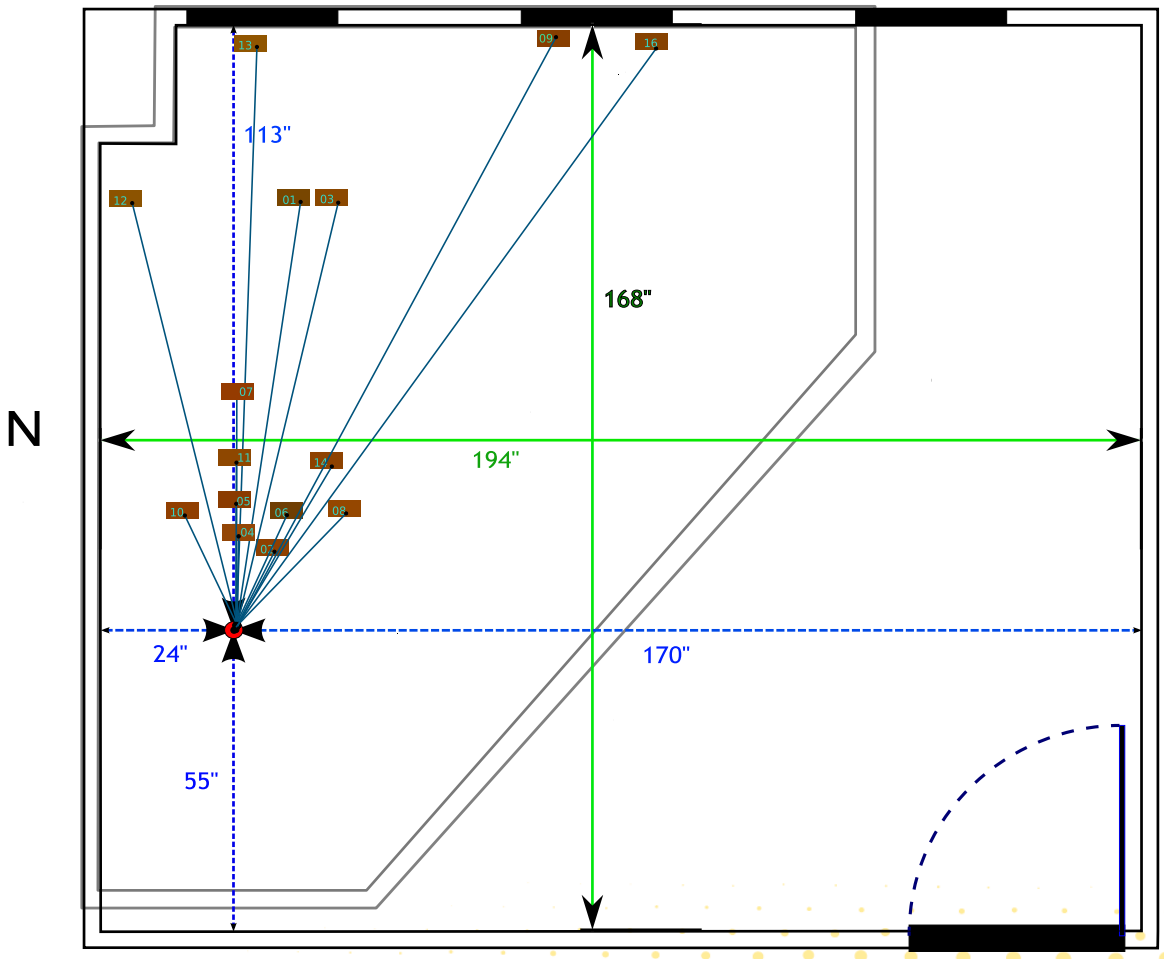
- Collect speech samples from 600 new participants
- Three (3) on-site sessions recorded via a cross-channel collection platform equipped with 15 distinct microphones
- Two recording rooms with different characteristics (size, shape)
- Duplicate methods, recording equipment, microphones, and mic distances in both rooms
- Each subject used only one interview room (3 times)
- Sixteen (16) telephone calls recorded via the LDC telephone collection platform (robot operator)
  - Three (3) telephone calls recorded via the cross-channel collection platform during on-site sessions (one per session)
  - Thirteen (13) telephone calls recorded outside of the LDC

Linguistic Data Consortium

Recruitment

- Subjects expected to participate in both on-site sessions at LDC and telephones calls conducted both at LDC and elsewhere
- Study required subjects to make three visits to LDC (all recruiting was done within the Philadelphia area)
- Recruitment was conducted via web advertising, flyers, and word of mouth
- Participants must meet the following criteria:
  - Native speakers of American English
  - First time participant
  - Over 18 years of age
- Initial enrolment with contact information via phone and/or email
- Demographic data collected in person at 1st in-house session
- Unique Personal Identification Number issued to each participant
- LDC recruited 748 participants to allow for natural attrition

Room Layouts: same mics, different walls



On-Site Sessions

- Goal: engage the subject in conversation
- Subject converses face to face with the interviewer, also reads text prompts
- Interviews were intended to be “sociolinguistic” style: minimizing interviewer speech, maximizing speech from the subject, reducing formality (to contrast with read-speech style later in the session), involvement with topics, more vernacular...
- Sessions Structure:
- Repeating questions: 1 min.
  - 6 short questions
- Informal conversation: 14 min.
  - Interviewer guided the conversation by taking notice of the participant’s interests
- Transcript reading: 15 min.
  - Participant read through a list of 335 utterances
  - Reading stopped after 15 min. whether or not the participant finished the list
    - Participants were asked to start over from the top if the entire list was read under 15 min.
- Telephone call: 10 min.
  - First session: High Vocal Effort (HVE)
  - Second session: Low Vocal Effort (LVE)
  - Third session: Cell phone
    - Plus an optional cell phone call, made outdoors after the interview session (few people did this)

In-Session Telephone Calls

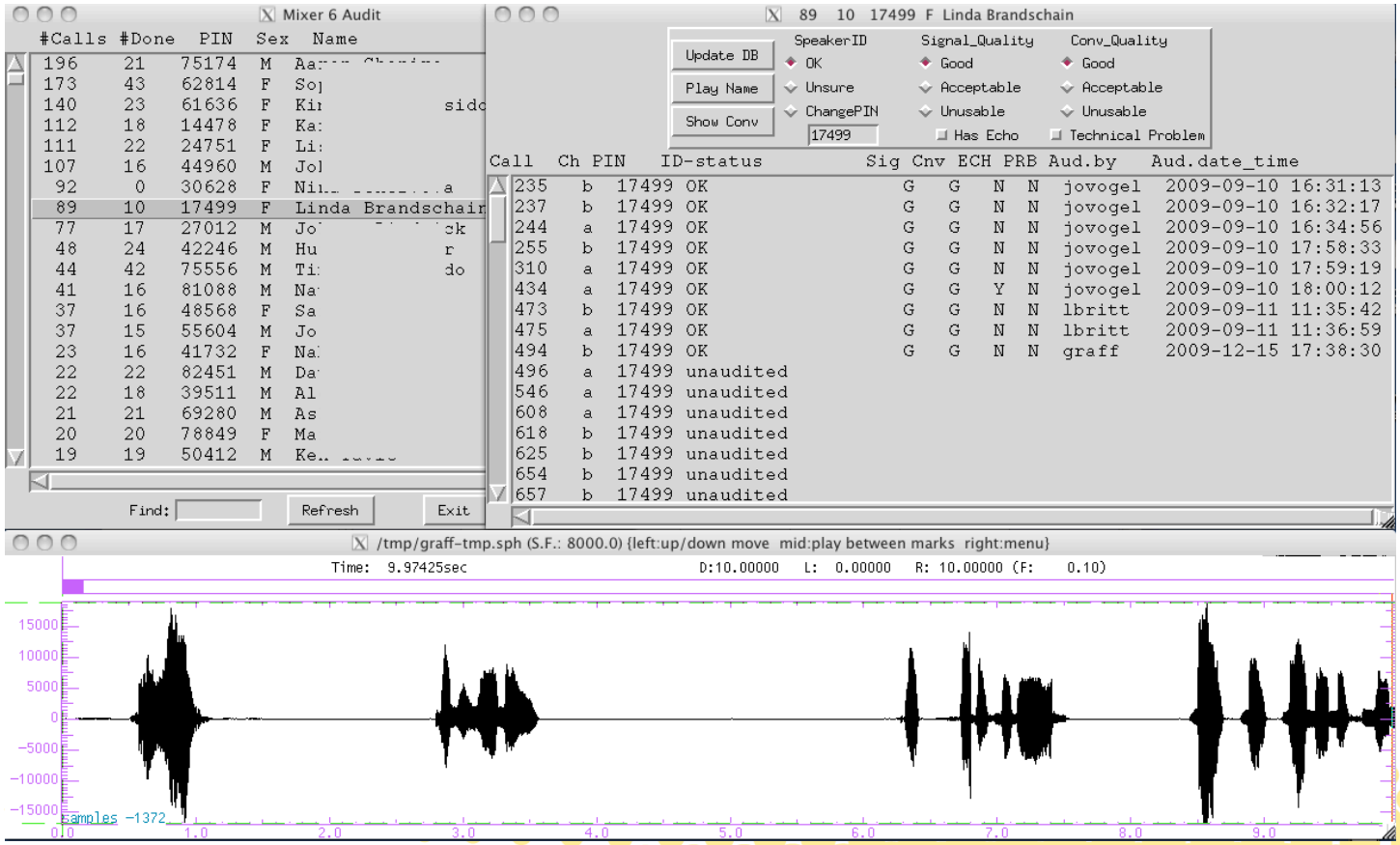
- All on-site calls were coordinated with LDC staff as confederates to guarantee bridging through the robot operator, and assure consistent interlocutor behaviour for LVE and HVE calls.
- High Vocal Effort
  - Brown noise was generated and fed to the participant only, through isolating headset earphones (interfering with the interlocutor's voice)
- Low Vocal Effort
  - The participant's voice and interlocutor's voice were both amplified in the headset earphones
- Cell Phone Call
  - LDC used two different cell phones using two different cellular networks and network types
  - Cell phones themselves are similar in type to what is currently popular in the marketplace and provided adequate representation of the differences between phones and networks to allow for comparisons

Pre-audit audio processing

- CTS:
  - robot operator creates pairs of raw 1-ch 8-kHz u-law sample files
  - daily uploads are copied to 2-channel SPHERE u-law files
  - VAD (esps-based) creates "transcript" (time stamps only)
  - If 5+ minutes of speech, queue file for manual audit; otherwise, mark file as "too short" (if subjects complain, review manually)
  - Manual speaker-ID audit done within days after recording
- Interview sessions:
  - Xchan platform creates sets of mswav 2-ch "16-kHz" pcm files (actual sample rate found to be 15899 Hz)
  - daily uploads are demuxed, resampled, and flac-compressed
  - Manual audits done at end of collection.

Telephone Audit Tool (first used in Mixer-3)

- Subjects sorted by # of calls recorded, list includes audit progress, gender, name
- All call sides are shown/accessible in call-detail window for a chosen subject
- Auditor can display calls for 2nd subject in 2nd call-detail window, if desired
- Mark PIN as incorrect if necessary, enter correct PIN if possible



Linguistic Data Consortium

Interaction with NIST

- NIST personnel (Greenberg, Fiscus) have login access on LDC servers, can view/copy corpus data as needed
- Special Mysqsl database account for NIST user, granting access to designated tables and fields (no access to subjects' personal identifying information)
- LDC compiles standardized metadata tables at completion of project/collection: subjects (IDs + demographics only), calls (subj_ids, "encrypted" phone numbers, audit results), and interviews (subj_ids, session info, audit results)
- Usable corpus inventory is "locked in" based on contents of final metadata tables.

Linguistic Data Consortium

Mixer 6 Delivery to NIST

- 3805 telephone conversations (incl. vocal effort calls) collected between July 2009 and Jan. 2010
- 1299 complete interview sessions (conversation + reading + phone call in a single audio recording)
- 584 subject-IDs represented (incl. LDC staff)

Calls / IVs	0	1	2	3	Total
0-7	38	78	40	22	178
8-15	8	13	21	62	104
16	7	7	37	251	302
Total	53	98	98	335	584

Greybeard Collection

- Goal: create a corpus that permits longitudinal study of the effect of aging on speaker recognition performance
- Collect conversational telephone speech from subjects who had participated in previous studies published by LDC
- Greybeard participants had to have at least 5 calls recorded in earlier studies (calls had to be at least two years old prior to the beginning of the Greybeard collection)
- Previous conversational telephone speech collections relevant to speaker recognition include:

  Switchboard 1

  Switchboard 2, three phases

  Switchboard Cellular, two phases

  Mixer Series

Greybeard Recruitment

- Based on experience from previous collections, LDC over-recruited and set the participants' goals higher than the research needs.
- Stated goal was to get 10 new calls from each of 100 speakers who had 5+ calls from earlier collections.
- We identified 206 candidates, sometimes using web searches for current contact information.
- We set participant payment incentives based on completing 12 calls
  - A subset of 25 participants was asked to complete 24 calls to assure a yield of 20 participants completing at least 20 calls

Linguistic Data Consortium

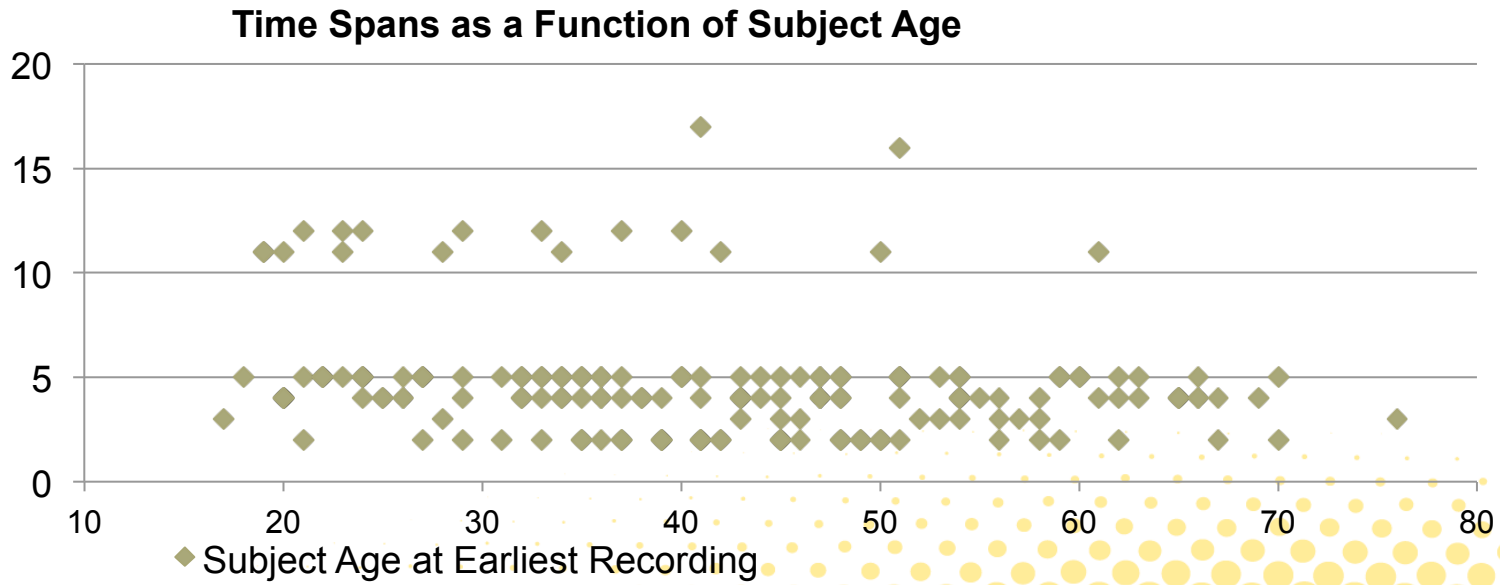
Greybeard Corpus Creation

- Call collection was rapid: 5 weeks in Oct-Nov. 2008
  - Normal Mixer3-style audit was done on these calls
- Metadata from legacy collections (pre-Mixer3), originally created in separate databases, were collated and conditioned for inclusion in our current "telco" master collection database:
  - Single, global "subjects" table with unique numeric SUBJ_ID, contact info and demographics for all subjects in all collections
  - Separate / parallel "subj", "calls" and "audit" tables for each collection project, mapping project-specific PIN to global SUBJ_ID
- All legacy calls were identified for all Greybeard subjects
- Full speaker-ID audit covered Greybeard and legacy calls.

Linguistic Data Consortium

Greybeard Delivery to NIST

Corpus Name	Collection Epoch	M	F	Total	Calls
Switchboard-1	1991-1992	2	0	2	36
Switchboard-2	1996-1997	14	2	16	362
Mixer 1 and 2	2003-2005	36	66	102	2358
Mixer 3 (1st phase)	2006	21	30	51	828
Greybeard	2008	71	95	166	1097



SRE08 Data (from Mixer-3/4/5 collections)

- SRE08 test segments came from:
  - 8446 CTS recordings collected over 2 years (12/2005 - 12/2007)
    - Over 1300 subjects in three collection projects combined
    - ~2000 Mixer-3 calls using ~20 languages other than English
    - ~890 Mixer-3 subjects providing both English and non-English speech
  - 1013 Multi-channel interview sessions collected in 2007 (Mixer-5)
  - 127 Multi-channel phone conversations collected in 2007 (Mixer-4)
- Synopsis of Mixer-3:
  - 14-month collection, many subjects, calls and languages
  - Designed to supply data for SRE06, LRE07 and SRE08
  - Web recruiting only (minimal staff interaction with participants)
  - Spkr-ID audits on all calls, plus lang-ID audit on LRE calls only

SRE08 Data (from Mixer-3/4/5 collections)

- Synopsis of Mixer-4:
  - 7-month collection, English only
  - 135 subjects: in-house multi-channel CTS recordings at LDC,ICSI
  - 143 subjects recruited via web (Mixer-3 style, minimal interaction)
  - Total of 256 multi-channel CTS sessions (247 used in SRE08)
- Synopsis of Mixer-5:
  - Ran concurrently with Mixer-4, English only
  - 340 subjects: in-house multi-channel interviews at LDC,ICSI
  - Up to six 30-minute M-C sessions (3 hrs) on 3 different days
- Both collections: up to 12 normal CTS calls per subject
  - No separation of subject pools in normal call collection

Problems found in SRE08 (1)

Problem 1: Some subjects were mislabeled as to gender

- Scope of problem: ~12 speakers mislabeled – 1%
- Initial Cause: Web enrollment, where demographic info is self-reported (not vetted by LDC staff at enrollment time)
- Aggravating factor: early version of web enrollment form may have provided a "default" answer for "gender" question, rather than requiring an explicit input from participants.
- Additional factor: speaker-ID audit interface has no direct method for updating speaker demographics (incl. gender) – primary focus is on speaker-ID and call quality.
  - Auditor must first notice gender discrepancy, then use a separate tool to update gender field in global "subjects" table

Problems found in SRE08 (2)

Problem 2: One SUBJ_ID number associated with two voices

- Scope of problem: 2 subj_ids each involved data from two distinct speakers (i.e. 2 ID#s, 4 voices) – 0.15%
- Initial cause: manual error by LDC auditors.
- One case: an innocent mistake by normal participant (used wrong PIN), then a procedural mistake by an auditor (only listened to recent calls for this subj_id, but was supposed to listen to previous calls also, even though those had already been audited)
- Other case: subj_id was used by multiple LDC staff, auditor mistakenly assumed that audit decisions "wouldn't be used"

Problems found in SRE08 (3)

Problem 3: A single voice appearing with multiple SUBJ_IDs

- Scope of problem: 10 individuals using 14 subj_ids – 1%
- Initial cause: An individual decides to cheat in order to get paid more money, and figures out ways to enroll multiple times via the web recruiting form (defeating normal checks based on name, email address, etc).
- Aggravating factor: Although the speaker-ID audit tool allows comparing two subj_ids at once, this is complicated and difficult for auditors; search techniques are limited.
- 153 duplicate enrollments caught in 2006-2007 (usually when preparing payments), but we missed 14 subj_ids

Linguistic Data Consortium

Factors in error reduction

- Face-to-face or direct-contact recruiting (don't believe what comes from the web)
- Collaboration between LDC staff and available SR technology to do focused searches
  - Acquire and use one or more SR systems in-house
  - Closely monitor subjects who use same address/phone number
  - "Adjudicate" NIST SRE system results before releasing final scores
- Incremental improvement/expansion of audit procedures

Issues with new data in SRE10

- Multiple-enrollment NOT an issue: all new participants were carefully recruited and screened (Greybeard and Mixer-6)
  - LDC staff using multiple PINs still happens, but easy to spot and fix
- Most "empty segments" in train/test data came from Greybeard
  - Problem: results of global audit were kept in one "cross-project" table, but call inventory metadata for NIST was drawn from separate "project-specific" tables – recent corrections were omitted from initial delivery, so...
  - Test set was built from faulty table data, but...
  - Corrected table data was provided before release of final scores.

Mixer-6/SRE10 Multi-mic audio

- Interview sessions:
  - Xchan platform creates sets of mswav 2-ch “16-kHz” pcm files (actual sample rate found to be 15899 Hz)
  - daily uploads are demuxed, resampled, and flac-compressed
  - Manual audits done at end of collection.

Inter-auditor agreement

- Purpose: re-do all audit decisions for Mixer-6 calls
- ~60% of 7500 sides (re)done as of Thursday a.m. (6/4)
- Mostly by different auditors (some decisions by same auditor, but months after first pass)
- No knowledge of previous decisions
- Replicates initial audit *almost* identically (but using output of pass-1 as input to pass-2)
- For "signal" and "conversation" quality, important distinction is between "Unacceptable" vs. "Good or Acceptable"
  - How useful / informative is "Good" vs. "Acceptable"? Keep that?

Inter-Auditor Agreement

- Signal Quality (Good / Acceptable / Unusable)

	G	A	U	(nd)	Total
G	2058	465	7	0	2530
A	1170	454	20	1	1645
U	7	5	12	0	24
(nd)	0	3	1	0	4
Total	3235	927	40	1	4203

- Conversation Quality (Good / Acceptable / Unusable)

	G	A	U	(nd)	Total
G	3714	190	15	3	3922
A	178	65	6	0	249
U	6	5	18	0	29
(nd)	3	0	0	0	3
Total	3901	260	39	3	4203

Inter-Auditor Agreement

- Echo (y/n)

	Y	N	(nd)	Total
Y	16	140	0	156
N	41	4001	4	4046
(nd)	0	1	0	1
Total	57	4142	4	4203

- "Tech.Problem" (y/n)

	Y	N	(nd)	Total
Y	12	13	0	25
N	21	4152	4	4177
(nd)	0	1	0	1
Total	33	4166	4	4203

- Speaker-ID:
  - 35 call-sides: 1st auditor said "OK", 2nd said "unsure"
  - 3 call-sides: 1st auditor said "unsure", 2nd said "OK"
  - 1 call-side: 1st auditor said "OK", 2nd set a different subject-ID

Conclusions/Questions

- Three classes of agreement/disagreement:
  - PIN disagreement: 0.02% (1 out of 4200)
  - sig, cnv, tech.problem: <1% each
  - echo: 4%
- Remaining questions: Given this level of agreement, should we:
  - keep doing it this way?
  - strive to improve training and consistency?
  - stop asking for these decisions?