Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora*

Christopher Cieri, Linda Corson, David Graff, Kevin Walker
{ccieri|corsonl|graff|walkerk}@ldc.upenn.edu
Linguistic Data Consortium, 3600 Market Street, Philadelphia, PA 19104

*Parts of this work were supported by funding from the Federal Bureau of Investigation, the Department of Defense and the Intelligence Technology Innovation Center under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Acknowledgements

- Thanks to the following who have supported the Mixer projects via sponsorship and/or consultation.
  - Walt Andrews (DoD)
  - Joe Campbell (MIT-LL)
  - George Doddington (SRI)
  - Jack Godfrey (DoD)
  - Fred Goodman (MITRE)
  - Audrey Le (NIST)
  - Mike King (ITIC)
  - Tina Kohler (DoD)
  - Alvin Martin (NIST)
  - Nikki Mirghafori (ICSI)
  - Nelson Morgan (ICSI)
  - Hirotaka Nakasone (FBI)
  - Barbara Peskin (ICSI)
  - Joe Picone (ISIP)
  - Mark Przybocki (NIST)
  - Doug Reynolds (MIT-LL)
  - Reva Schwartz (USSS)
  - Wade Shen (MIT-LL)

SRE Data

- **Some properties of robust Speaker Recognition systems**
  - text independence
  - channel independence
  - language independence
- **Data for system development and evaluation should support those requirements**
  - multiple, variable samples per speaker
    - » generally: conversational speech with the topic varying
    - » more recently: increased variation in speech genre
  - collection channels also vary across or even within sessions
    - » generally: subjects use multiple telephone handsets
    - » more recently: some sessions recorded via many channels
  - multiple languages sampled
    - » generally: multiple collections in different languages
    - » more recently: collections in which bilingual subjects use at least two target languages, one per session

Collection Protocol

- Switchboard
  - each speaker makes multiple calls
    - subject initiates call, robot operator calls other subjects to find match meeting specific criteria
      - pair has not spoken before, both interested in same topic
  - brief: six-minutes in duration
  - conversation among strangers
  - using assigned topics
  - collected as 4-wire data
- Mixer Enhancements
  - new protocol adapted to today’s telephone use where
    - voice mail, call screening, call forwarding
  - such that
    - robot operator calls all available subjects at times they specify
    - subjects also permitted to call robot operator
    - constraints lifted, all pairings allowed
  - multiple languages collected using bilingual speakers
    - robot gives priority to speakers of same native language
    - some hours/days were devoted to non-English calls
  - intensively cross-channel
    - multichannel interface, recording application, 8 or 14 sensors
    - calls collected by robot operator simultaneously
    - deployed cross channel recording system at multiple sites
  - compensation = core fee + special features + completion bonuses

Comparison of Phases

	SB	M1	M2	M3	M4	M5
Core Calls (8+)	✓	✓		✓	✓	✓
Variable Environments	✓					
Unique Handset (4+)	✓	✓	✓	✓	☒	✓
Extended Data (20+)		✓	✓	✓	✓	
Multilingual (4+)		✓		✓	☒	
Cross Channel (2 or 4)		✓	✓		✓	
Transcript Reading (2+)		✓				✓
Interviews (6)						✓

Mixer 3 Plan

- Data for development and evaluation of Speaker Recognition systems
- Data for development and evaluation of Language Recognition systems
  - CallFriend-2 protocol
    - subjects complete single call to friend/family
    - within the continental United States or Canada
    - topics of their choosing
    - call was toll-free up to 30 minutes, both caller and callee were compensated
  - worked well through the 1990's
    - more than 1000 calls
    - more than a dozen linguistic varieties including: American English, Canadian French, Egyptian Arabic, Farsi, German, Hindi, Japanese, Korean, Mandarin, Russian, Spanish, Tamil and Vietnamese (all in LDC Catalog)
  - New collection too slow presumably due to lack of incentives
    - free phone call worth less than it used to be
    - 1 USD per minute is good on average but 1 USD/minute * 10 minutes = $10 (only)
- Mixer 3 could meet both needs
  - bimodal distribution of speakers with respect to the number of calls completed
    - many complete 0 calls or 1 call before dropping out
    - of remainder approximately 70% accomplish 80% of the established goals
  - With goals and compensation set carefully,
    - subjects making 1 call provide data for LRE
    - subjects making target number provide 1 calls for LRE plus remainder for SRE
  - To ensure robust evaluation
    - calls used for the first evaluation not released until the second evaluation complete

Mixer 3 Outcome

- Mixer 3 performed roughly as expected
  - actually outperformed expectations for SRE but fell short for LRE
- Where CallFriend generated
  - few calls
  - most of which were useful for LRE
- Mixer generated
  - large number of calls
  - most of which were useful for SRE
  - smaller percentage useful for LRE
- Specifically
  - >2900 Mixer 3 subjects each made a call in one of
  - 32 languages including Aceh, Amharic, Bengali, Burmese, Chechen, 4 dialects of Chinese, 3 dialects of English, Farsi, Georgian, Guarani, Hindi, Italian, Japanese, Khmer, Korean, Lao, Punjabi, Russian, Spanish, Tagalog, Tamil, Thai, Tigrigna, Urdu, Uzbek, Vietnamese
- For SRE
  - 19,951 calls
  - >1500 subjects completed 15 or more calls (compare to 400-600 in previous studies)
- However for LRE
  - distribution of calls across languages was uneven
  - have not yet reached goal of 100 calls in each language
  - some languages are poorly represented

Mixer 4 Plan

- Original plan to increase supply of both LRE and SRE data by collecting data from
  - 400 subjects who each make 10 calls in 4 new languages: Maghrebi Arabic, Hindu/Urdu, Korean, Tagalog
  - 100 subjects who make 20 or more calls
  - 200 subjects who make 4 calls from one of the project's multi-channel recording systems
  - 100 speakers who make calls from at least 4 unique handsets
- However, responding to the need for
  - more SRE data including
  - data from native speakers of English to support use of high level features
- The current plan for Mixer 4 is to include
  - 400 subjects who each make 10 calls in English
  - 100 subjects who make 20 or more calls
  - 200 subjects who make 2 calls from one of the project's multi-channel recording systems
- Additional LRE data will be collected via, claques, native speakers of a target language, who use their social networks to stimulate calling in those languages.
- LDC has recently used this method to reach targets for a number of languages that had fallen short under the CallFriend 2 and Mixer protocols

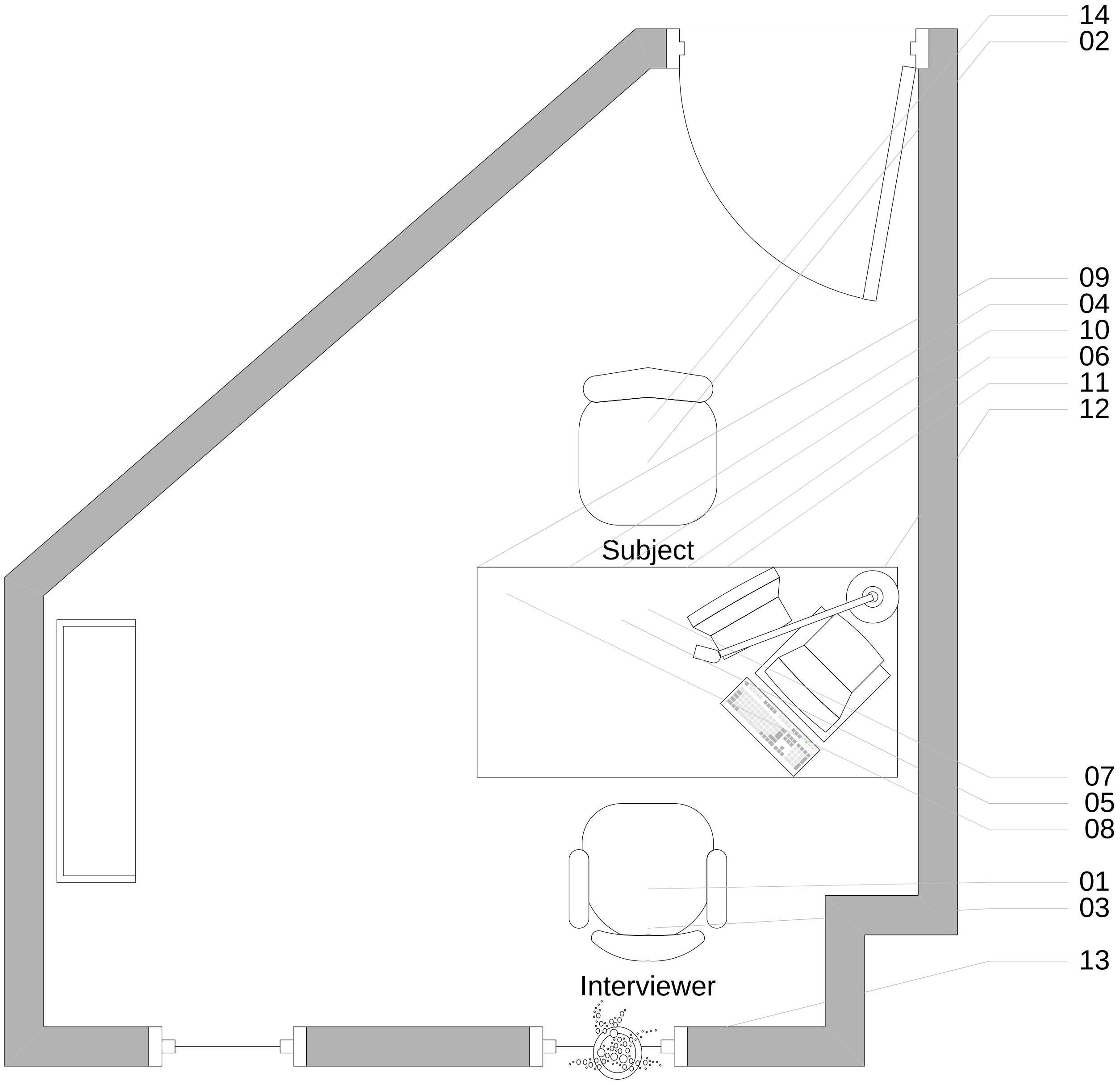
Mixer 5 Plan

- Based on feedback from Fred Goodman (MITRE), Mike King (ITIC), Jack Godfrey (DoD) and George Doddington (SRI/NIST), LDC made numerous changes to the Mixer protocol for Phase 5
- Cross-Channel collection system rebuilt
  - Several microphone used in Mixer 1 & 2 cross-channel have been replaced.
  - Several new microphones have been added.
  - Recording system upgraded to handle 16 channels (was 8)
  - Same system will be used in Mixer 4
- 10 telephone conversations augmented with 6 interview sessions.
- Interview sessions collected at LDC and ICSI.

Sensors in Cross Channel Sessions

#	Microphone	Placement
01	Shure MX185 Lavalier	Worn: Interviewer’s clothing under chin.
02	Shure MX185 Lavalier	Worn: Subject’s clothing under chin.
03	Etymotic Link-It microarray	Worn: Interviewer’s ear.
04	Shure MX418S Podium	Fixed: Desk Front, Subject's Center
05	Crown PZM-6D	Fixed: Desk Top, Subject's Center
06	Audio Technica AT3035	Fixed: Desk Front, Subject's Right
07	Audio Technica Pro45	Fixed: Hanging, Subject's Center
08	Panasonic Camcorder	Fixed: Desk Top, Subject's Right
09	R0DE NT6	Fixed: Desk Front, Subject's Far Left
10	R0DE NT6	Fixed: Desk Front, Subject's Left
11	R0DE NT6	Fixed: Desk Front, Subject's Center
12	R0DE NT6	Fixed: Desk Front, Subject's Right
13	AcoustiMagic Array	Fixed: Wall Mounted, Subject's Center
14	Lightspeed XLC-20	Worn: Head Mounted, Only During Calls

Cross Channel Interview Room

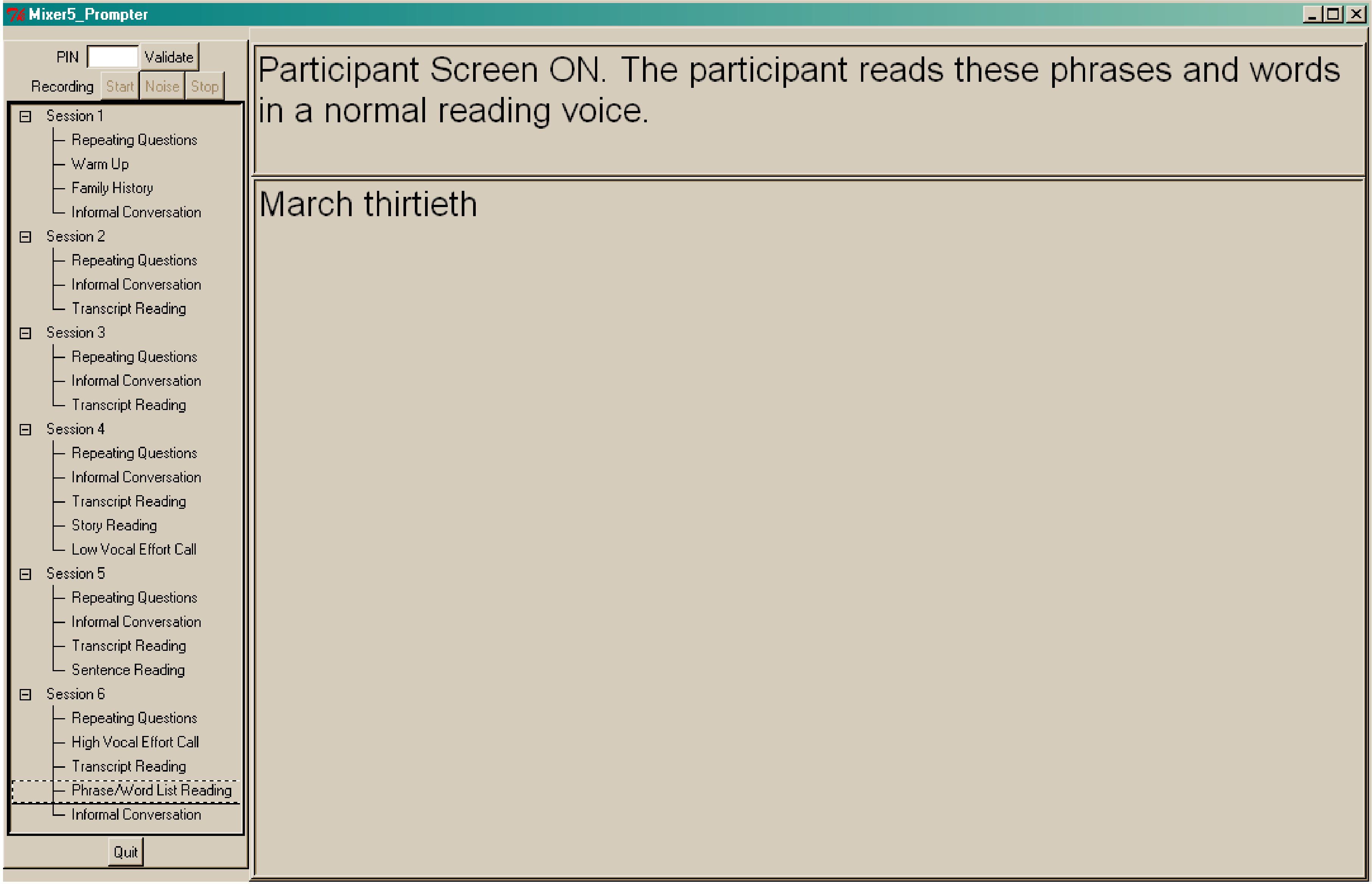


Cross Channel Recording Room

Mixer 5 Interviews

- Repeating Questions: to elicit multiple instances of speech where the same words appear each session begins the same simple questions
- Family and Personal History: to elicit demographic information, interviewer focuses on the personal and family history of the subject
- Informal Conversation: to elicit informal, speech in which the subject's attention is directed toward the topic under discussion and away from the form of language used, interviewer ask questions designed to identify subject's interests
- Transcript Reading; subject reads transcripts of utterances from previous phone conversations that have been edited to remove disfluencies & sorted by association with a topic
- Story Reading: subject read phonetically rich stories
- Sentence Reading: subject reads subset of TIMIT sentences selected to elicit features that distinguish dialects of American English
- Low/High Vocal Effort Speech: subject conducts phone conversation while wearing sound isolating headphones where signal, side tone and noise levels are modified to encourage changes in vocal effort
- Phrase/Word List Reading: subject reads word and phrase lists containing items that are diagnostic for American English dialects.

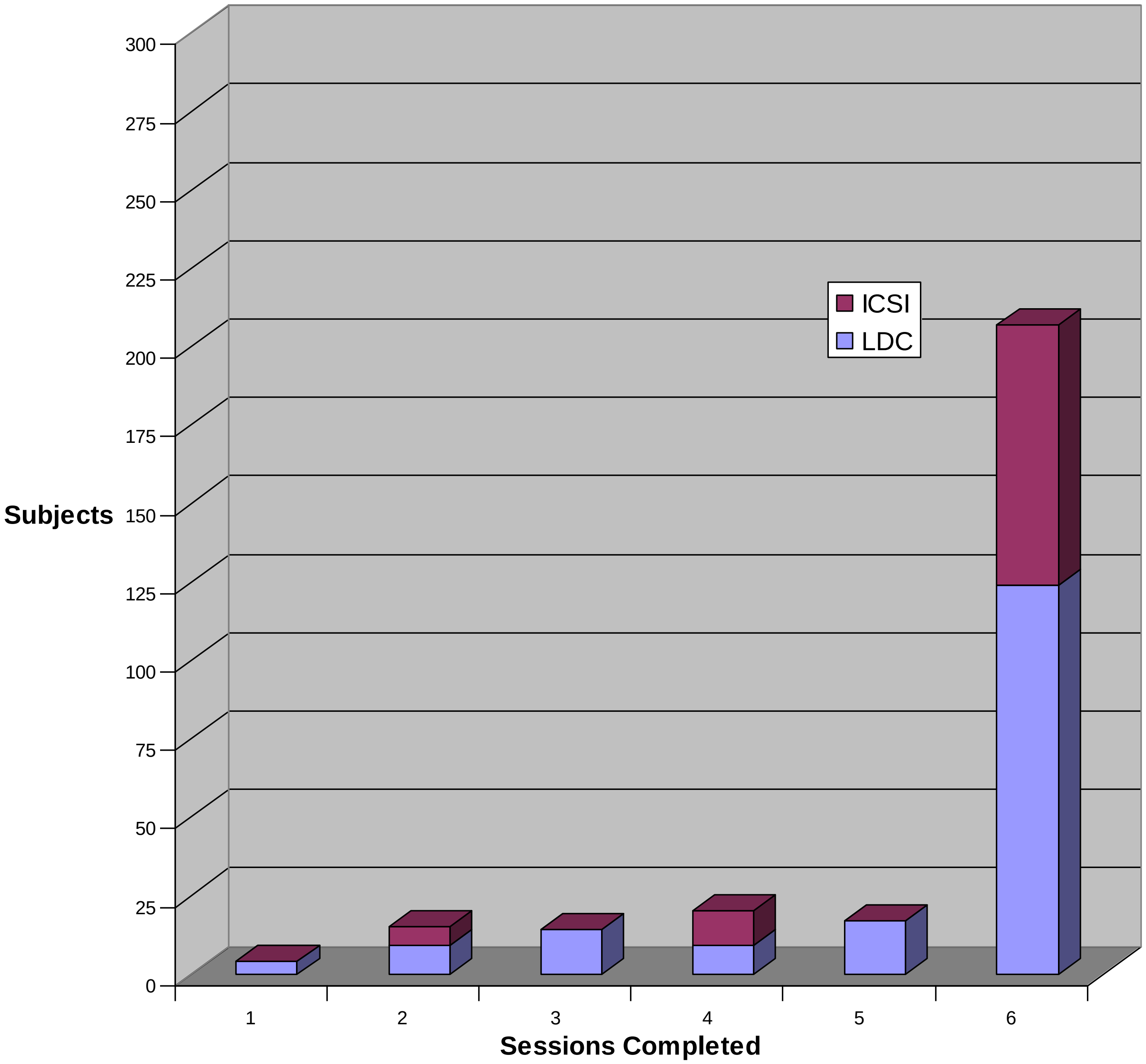
Mixer 5 Prompter



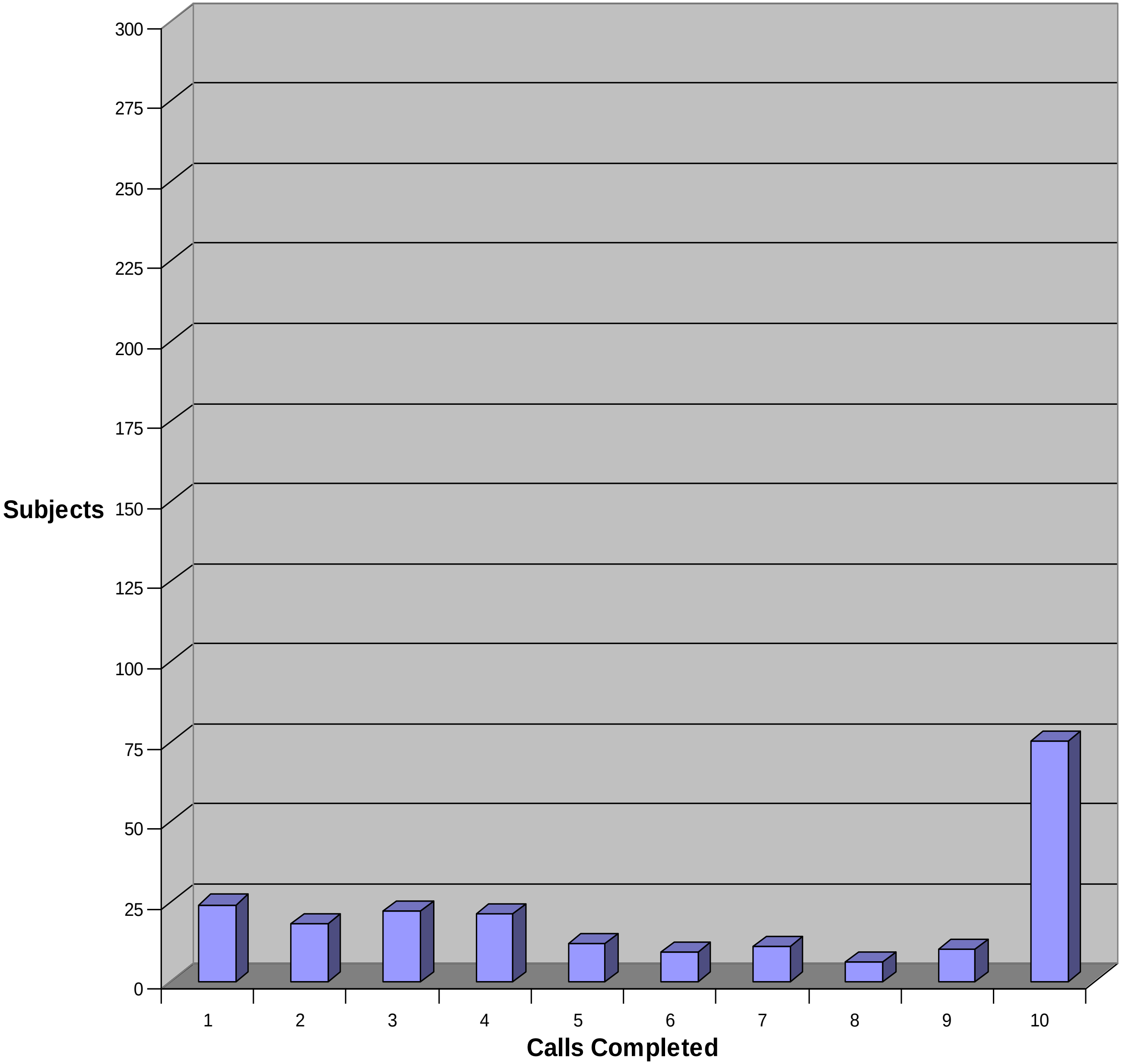
Participant Screen ON. The participant reads these phrases and words in a normal reading voice.

March thirtieth

Mixer 5 Interview Progress



Mixer 5 Call Progress



Future Work

- **Mixer 1 & 2**
  - in LDC publication pipeline
- **Mixer 3**
  - used in SRE 2006 & LRE 2007
  - remainder could be used in SRE 2008
- **Mixer 4**
  - collection underway
  - will be used in SRE 2008
- **Mixer 5**
  - interview collection ahead of schedule
  - phone call collection also well underway
  - may be used in SRE 2008, may be used for new SRE program?
- **LDC would like to**
  - conduct a longitudinal study using subjects from previous 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.**