The Mixer 6 Corpus:
Resources for Cross-Channel and Text Independent Speaker Recognition

Linda Brandschain, David Graff, Christopher Cieri, Kevin Walker, Chris Caruso, Abby Neely
Linguistic Data Consortium
University of Pennsylvania
E-mail: {brndschn,graff,ccieri,walkerk,carusocr,aneely}@ldc.upenn.edu

Abstract

The Linguistic Data Consortium's Human Subjects Data Collection lab conducts cross-channel speech collections to develop corpora for use in speech, speaker and language recognition research and evaluations. The Mixer collections have evolved over the years to best accommodate the ever-changing needs of the speaker recognition research community and to hopefully keep one step ahead by providing increasingly challenging data. Over the years Mixer collections have grown to include socio-linguistic interviews, a wide variety of telephone conditions and multiple languages, recording conditions, channels and speech acts. This paper describes Mixer 6, the most recently completed collection, whose object was to record speech via a variety of channel types and in a variety of situations that vary formality and model multiple naturally occurring interactions.

1. Goal of this collection

Mixer 6 is the latest in a series of collection designed to provide language resources for speaker and language recognition research (Brandschain, et. al., 2008, Cieri, et. al., 2006). The goal of the Mixer 6 project was to provide new and challenging data for the development and evaluation of both text independent and text dependent speaker recognitions technologies. This data was to exercise the technologies' ability to deal with variations in time, channel and interactional situation. The specific tasking was to collect speech samples from 600 new participants, who would complete multiple on-site sessions recorded via a cross-channel collection platform equipped with 15 distinct microphones and multiple telephone calls recorded via the LDC telephone collection platform (robot operator) and, at least for a subset via the cross channel platform. A subset of this data was intended for use in the 2010 Speaker Recognition Evaluation (SRE) campaign organized by the U.S. National institute of Standards and Technologies (NIST 2010).

2. Recruitment

For this project, LDC recruited 748 participants to allow for natural attrition. All were native speakers of US English. At the sponsors' request, the subject pool consisted only of subjects who had never participated in a previous LDC speech study. The subjects were expected to participate in both on-site sessions at LDC and telephones calls conducted both at LDC and elsewhere. Since this 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, the last proving again to be a highly efficient, low cost and productive method.

3. Registration and Scheduling

To register, participants called the toll-free number assigned to the project between the hours of 9am and 6pm Monday through Friday. Project staff on hand to take the calls, explained the tasks, answered any questions and collected the necessary subject information for entry into a database. Once subjects were registered they were asked to schedule the first of their onsite interviews. Subject had to demonstrate commitment to the study by completing some interviews before they would be permitted to complete their telephone calls.

4. Interview Protocol

4.1 Interview

Participating in each interview were a subject, and an interviewer. The interviewer's goal was to engage the subject in conversation and guide the subject through a series of more formal speech elicitation exercises. The subject was required to converse, face to face, with the interviewer and read prompted text. Interviewers were trained to ask questions and conduct the interview in a manner that would minimize their own verbal input and maximize the speech from the subject.

4.2 Telephone Calls

Telephone calls made onsite consisted of the subject, an interviewer and a confederate. The confederate was a staff member whose responsibilities included participating in these

calls. The subject, assisted by the interviewer, engaged in calls in which the elicited speech was characterized by high and low vocal effort as discussed below.

5. Description of Tasks

5.1 Repeating Questions

To provide a small amount of data to support investigations into variability within the speech of a single subject, all were asked the same set of short questions at the beginning of each session. Subjects were informed that they would answer these questions each time. Because the questions were the same, the answers were generally the same in content though not identical in form across sessions for a single subject. The questions and the answers they elicited took approximately one minute. During this time the subject could also get comfortably settled in the chair and the interviewer could check the sound levels and equipment to make sure that all was working properly prior to beginning the Informal Speech section of the session.

5.2 Informal Speech

The goal of the Mixer 6 interview sessions was to record speech that varied in formality and modelled multiple naturally occurring interactions. Structurally the sessions began with informal interviews and closed with the more formal elicitations of read speech. The goal of the former was to elicit informal, speech in which the subject's attention was directed toward the topic under discussion and away from the form of language used thus increasing the probability that the subject's speech approximated vernacular. The more formal elicitations were intended to elicit speech that was both phonetically rich and focused upon specific linguistic phenomena.

The profile of the sessions was expected to be formal at the beginning of the first session with formality generally decreasing over time but with style shifts during the prompted speech exercises.

The interviewer led the subject through the informal sessions by asking series of questions taken from, adapted from or inspired by the Interview Modules introduced by Labov and his colleagues and commonly used in socio-linguistic research (Labov 1984). Indeed, the entire methodology of the Informal Speech section is adopted wholesale from the best practices of sociolinguists. This portion of the session took 14 minutes to complete. Although the order and structure of the sessions was fixed, the order of modules within the Informal Speech sections was not. The modules have been written so that the interviewer can sensibly begin from the top and progress straight through; however, this was not required or even encouraged. The modules form a network that the interviewer and subject traverse in different ways depending upon the age, sex and interests of the latter. At the beginning of each line of questioning, the interviewer watched for signs of interest on the part of the speaker: posture, eye movement and volume of speech and nature of response. The interviewer pursued topics that interested the subject and abandoned topics that produced little or no response or else produced signs of uneasiness or boredom. The interviewer tried to make natural transitions from one module to the next and promote the flow of conversation by maximizing shared knowledge, minimizing authority and minimizing consequences (Labov 1984). If the subject mentioned an interest, the interviewer could move directly to a relevant module or else improvise questions.

The interviewer's questions were brief (taking no more than 5 seconds to deliver) and their contributions to the conversation, stories and opinions for example, were as short as possible while still being real contributions. The interviewer did not read from the question modules but rather became familiar with them and then asked the questions from memory as naturally as possible. Where appropriate the interviewer encouraged the subject to tell stories about events in the subject's past and to describe objects or procedures in detail.

5.3 Transcript Reading

To provide data to support research into phonetic factors in speaker recognition, LDC collected multiple repetitions from each speaker of a large set of utterances read from transcripts of prior telephone speech collections. From the tens of thousands of transcribed utterances, LDC culled a subset that were uttered originally with few disfluencies, that represented the broad range of topics included in prior collections, that made sense, that would be reasonable to read, and that a subject might be expected to utter naturally.

The resulting list consisted of 335 utterances including some containing just one or two words. The sentences appeared on a monitor that was facing the subject. Participants were asked to read the sentences as prompted in as natural a manner as possible. When they completed an utterance the interviewer changed the prompt. From our experience in previous studies we found that

we could eliminate lengthy instructions and that the subjects relax into this task rather quickly. It is our belief that instructions as to how the sentences should be read heighten the subject's awareness and produce unnatural speech. The subject was presented with the same list of sentences in the same order and began at the top of the list in each session. This section was 15 minutes in length. Some of the participants were able to complete the entire list of 335 sentences, within the 15 minutes allotted to this task, in which case they began again at the top. This overall approach guaranteed the greatest number of repetitions of the utterances.

5.4 Telephone Calls

This study also collected telephone calls from each participant. The goal was for each participant to complete a total of 16 telephone calls. All 16 were recorded via LDC's telephone collection platform. Participants were able to dial into our toll-free number or receive calls on a schedule they determined. When a participant initiated the call, the robot operator asked the caller to input their Personal Identification Number (PIN) and checked that PIN against a database of registered subjects. Callers also received instructions to input the number from which they were placing the call and asked to choose the type of phone being used: cell phone, land line, speaker phone, etc. Participants then heard a suggested topic for conversation. The robot operator then actively attempted to pair the participant with another on hold or placed a call to a participant who was listed as available to receive calls at that time.. The system checked each participant's PIN and prior conversation partners against other available callers and attempted to match those who had not previously conversed. Once paired, the calls were bridged and the subjects began the discussion. The topics, which changed daily, were designed to elicit conversation but there was no penalty for not discussing the proffered topic; subjects were free to discuss any topic upon which they and their call partner agreed. Each call lasted ten minutes.

A subset of 3 or 4 of the telephone calls was made while at LDC either prior to or just following an interview session. Of the subset, 3 calls were recorded in the LDC interview room and simultaneously recorded on the 15 channels. These calls followed specific guidelines and were designed to elicit differing levels of vocal effort. Two of these calls used a landline and the third a cell phone. The fourth on-site call was made via cell phone from the street corner just outside LDC. This spot was chosen as a typical, busy city corner with street, traffic and pedestrian noise occurring naturally. The forth call was optional for the participants. All on-site calls were coordinated with LDC staff as confederates. The two landline calls placed at LDC are designed to elicit either high vocal effort or low vocal effort from the participant.

To elicit high and low vocal effort calls the participants were asked to wear aviation grade headsets that isolated them from the sounds occurring in the room and were attached to a mixer that controlled the volume at which they heard their own voice, the voice of their conversation partner and optional noise. For low vocal effort calls, the subject's own voice was amplified encouraging them to speak more softly. To induce high vocal effort calls, subjects' own voices, and those of their partners, were attenuated and subjects were exposed brown noise at a decibel level high enough to make it difficult to hear themselves or their partners.

6. Cross-Channel Recording Equipment Specifications and Settings

The multi-track audio recording system used during cross-channel data collection consisted of the following equipment:

- Windows workstation
  - Intel Xeon dual-core 2Ghz CPU
  - ADS-Tech IEEE1394 adapter
  - 3Ware SATA adapter
  - 2GB RAM system memory
  - nVidia dual monitor video graphics adapter
  - two 19" LCD monitors
- one MOTU 896HD audio interface
  - eight analog audio inputs
  - mic level or line level
  - 0 to +40dB gain
  - optional phantom power
  - 8 line-level analog outputs
  - AES-EBU I/O
  - ADAT Lightpipe I/O
  - firewire connection to workstation
- one MOTU 8pre audio interface
  - 8 variable gain (0 to +40dB) microphone preamplifier inputs
  - connected to 896HD via ADAT lightpipe
- 1TB external RAID-1 hard disk drive connected to the workstation via eSATA
- one JBL audio monitor connected to the MOTU 896HD via AES-EBU

The system could be used to record up to 16 channels of audio from a mixture of line-level and mic-level inputs. Additionally, control tones, masking noises, and other audio sources

could be played back simultaneously via the MOTU 896HD audio outputs.

7. Microphones

The study used a total of 15 microphones or sensors of which one was devoted to the interviewer. Microphones were used and placed consistently in each interview session.

Interviewers were instructed to pay particular attention to the placement of two lavalieres, one mounted on their lapel the other on the subject's lapel, to facilitate diarization.

The microphone types used in this collection included: lavaliere, head-mounted, podium, PZM, studio, conference room, camcorder, shotgun, array and aviation headset. Some information about the microphones is withheld until the resulting corpus has been exposed in the relevant technology evaluations.

LDC conducted interviews in two separate recording rooms. Great care was taken to insure that microphone placements, distances, mountings and orientations were the same in both rooms. In addition, prior to the start of each session the room was calibrated and microphone levels checked by the interviewer.

8. Cell Phones

LDC used two different cell phones that, in turn, used two different cellular networks and network types. The 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. Information regarding phones and networks was available in the in the resulting data collected. Some information about the microphones is withheld until the resulting corpus has been exposed in the relevant technology evaluations.

9. Yield

The Mixer 6 collection yielded a total of more than 8700 calls and more than 1400 interview sessions. A large subset of these calls and interviews were delivered to NIST and the project sponsors and a smaller subset mixed with data from the Greybeard and prior Mixer collections and used in SRE10

	Calls			
Interviews	**0-7**	**8-15**	**>=16**	**Total**
0	215	4	1	**220**
1	55	12	3	**70**
2	16	11	35	**62**
3	20	20	368	**408**
Total	**306**	**47**	**407**	**760**

Table 1: Mixer 6 Subject pool by calls and interviews completed.

Tables 1 and 2 show, respectively the Mixer 6 subject pool by interviews and telephone calls completed and the structure and duration of the calls and interviews.

	Sessions			Phone Calls	Total Minutes
Activity	**1**	**2**	**3**		
Question	1	1	1		**3**
Conversation	14	14	14		**42**
Transcript Read.	15	15	15		**45**
High Vocal Effort	10				**10**
Low Vocal Effort		10			**10**
Cell , Indoors			10		**10**
Cell, Outside			10		**10**
Telephone Calls				120	**120**
	40	**40**	**50**	**120**	**250**

Table 2: Structure and duration of Mixer 6 calls and interviews.

10. Conclusion

Mixer 6 continued the trend of earlier studies providing speaker recognition data from a variety of speakers, sessions, interactional styles and microphone channels. Part of the resulting data is currently in use for SRE10. The corpus will be released generally after its use in open NIST evaluation campaigns.

11. References

Brandschain, Linda, Christopher Cieri, David Graff, Abby Neely, Kevin Walker (2008) Speaker Recognition: Building the Mixer 4 and 5 Corpora, LREC 2008, Marrakech, Morocco.

Cieri, Christopher, Walt Andrews, Joseph P. Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, Kevin Walker (2006) The Mixer and Transcript Reading Corpora: Resources for Multilingual, Cross-channel Speaker Recognition Research, LREC 2006: Fifth International Conference on Language Resources and Evaluation.

Labov, William, (1984) Field Methods of the Project on Linguistic Change and Variation. In Baugh, John and Joel Sherzer, *Language in Use: Readings in Sociolinguistics*, Prentice Hall.

NIST (2010), NIST Speaker Recognition Evaluation Web Page, National Institute of Standards and Technologies, Information Access Division, Multimodal Group, http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html.