# Greybeard – Voice and Aging

**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 Greybeard project was designed to enable research into the effects of aging on speaker recognition performance by providing data that had been collected over a long period of time. Since 1995, LDC has been collecting speech samples for use in human language technology research, development and evaluations, specifically to support speech, speaker and language recognition. By mining our earlier collections we assembled a list of subjects who had participated in these studies. The participants were then contacted and asked to take part in the Greybeard project. The only constraint was that the participants must have made numerous calls in prior studies and the calls had to be a minimum of two years old. The archived data was collected and collated by participant and subsequent calls were added to their files. This is the first longitudinal study associated with NIST SRE technology evaluations that we have been able to identify. The resulting corpus contains multiple calls for each participant that span time periods of two to 12 years in time. A subset of this data will be used in the NIST 2010 Speaker Recognition Evaluation (SRE) where it will enable speaker recognition researchers to explore the effects of aging on voice.

## 1. Introduction

Prior research has shown variation in voice as a function of aging. Biever and Bless's (1989) cross sectional study found greater shimmer, greater variability in mean airflow rate and more aperiodicity in their older female subjects. Decoster and Debruyneartin's (1990) longitudinal study found differences in fundamental frequency, the standard deviation of fundamental frequency and voice-onset time among 20 male news broadcasters reading identical text on two occasions with 30 years time intervening. Vipperla, Renals and Frankel (2008) found word error rates (WER) of their automatic speech recognition (ASR) generally greater for elderly subjects and indeed increasing with age. Although Yuan and Liberman (2008) achieved "near-100% text-independent identification accuracy on utterances that are longer than one second" using the same corpus, SCOTUS, the issue of speaker recognition performance where the target speaker has aged significantly between the time of enrolment and a given trial is relatively rarely trodden territory.

This goal of the Greybeard project was to create a corpus that permits longitudinal study of the effect of aging on speaker recognition performance by collecting conversational telephone speech from subjects who had participated in previous studies available to the Linguistic Data Consortium (LDC). To enable serious work on a challenging issue, Greybeard participants had to have completed at least 5 calls made in earlier studies and those calls had to be at least two years old prior to the beginning of the study.

The subset of previous conversational telephone speech collections relevant to speaker recognition includes the original Switchboard (Godfrey, et. al., 1992), the three phases of Switchboard II, two phases of Switchboard Cellular (Miller, et. al., 2001) and the Mixer series (Cieri, et. al., 2006, 2007). These previous collections share a number of features with Greybeard. From 200 to 600 subjects were recruited to complete from ten to 25 telephone conversations of three to ten minutes duration speaking to other participants, whom they typically did not know, about topics suggested a robot operator. The average duration of these studies runs into the months. Subjects are encouraged to complete a large number of calls from a single handset and then the remainder from multiple other handsets. In some studies, subjects were also required to change location so that some calls were conducted inside the home or office, some outside and some from within a moving vehicle where the caller was not the driver. In the later, Mixer, studies, bilingual subjects sometimes also spoke in languages other than English and sometimes conducted calls from within LDCs cross-channel rooms where they were simultaneously recorded on multiple microphone channels.

## 2. Recruitment

Knowing that subjects in telephone collection studies often drop out and that those who remain typically produce less data than requested of them, LDC over-recruited and set the participants' goals higher than the research needs.

In an effort to locate 100 appropriate speakers for the Greybeard project, LDC reviewed its historical records and identified, contacted and attempted to recruit 209 speakers that met the collection criteria. Participants were asked to complete 12 calls to assure a yield of at least 10 complete calls per subject. A subset, of 25 participants, was asked to complete 24 calls to assure a yield of 20 participants completing at least 20 calls.

### Finding Qualified Candidates for Participation

In order to identify participants, LDC programmers collated lists of participants from previous studies for which records were available who had made numerous

calls. These lists included all available contact information. Programmers also loaded information from earlier projects into our current participant database to create one universal subject database for LDC speech projects. Where possible, data collected from earlier databases were updated and each subject record was made available to project staff for search and edit via our subject management tool.

Project programmers then developed a list of qualified participants. The initial criteria for participation was that LDC have multiple previous recordings for each recruit and that the previous recordings be at least 18 months old prior to the beginning of the new collection. Native, male English speakers were preferred, and recruited first. However, previous experience has shown that telephone speech collections can fail simply because the initial subject pool is too small. When all potential males were recruited and the subject pool was still too small (<200), native, female English speakers were included. Non-Native English speakers were excluded from this collection.

Initially, a mass mailing was sent to the entire list of qualified male candidates. Returned emails from defunct email addresses were collected into a spreadsheet that also contained all available contact information including any phone numbers on record. Project staff then systematically reviewed the list and called each number, logging whether the former subject was interested in participating in Greybeard or not, or whether they failed to reach the subject or left a message. Naturally, the contact information for the more recent studies was more current than the older ones and the response and yield were both greater as well. After exhausting the phone lists, staff then attempted to locate former participants via the Internet and on sites such as Facebook and Myspace. For Switchboard, of the approximately twenty subjects that eventually joined the Greybeard study, approximately half came from mass mailing and a quarter each from calling and Internet research.

Initial recruitment among former male subjects yielded approximately 80 qualified males interested in participating. Since this was too small a pool with which to open the calling platform the study was then opened to qualified women. Our experience from previous studies showing that women are more receptive towards phone studies was repeated in Greybeard and the female pool filled up quickly. As the study progressed more qualified males were successfully recruited and the final numbers show an almost even 50/50 gender split for the study.

### Registration

Candidates registered via the Internet, by contacting LDC staff by phone or in person, at which time they were asked to provide demographic information to include with the research data. Personal identifying information is confidential and used for contact and payment purposes only. Such information is never shared with the research data. During registration, participants were also informed that they were free to leave the study at any time without penalty.

## 3. Data Collection

### Protocol

LDC followed the general outlines of previous call collection protocols in which a robot operator initiated calls to registered subjects at times and telephone numbers they specified and accepted calls from subjects pairing those who agreed to participate in a call at that moment. Due to the relatively small size of the Greybeard subject pool, LDC limited the number of hours during which the robot operator would be active in order to increase the probability of subjects connecting to each other. In addition, LDC staff was available to serve as conversation partners during the normal hours of operation, Monday through Friday, 9am to 6pm. The call sides from LDC staff were generally excluded from the database except in the case of a staff member meeting all of the criteria for the study.

### Operation

The LDC robot-operator was available daily from 2:00PM until 12:00 midnight EST with minimal down-time for maintenance. Subjects were asked to provide a schedule of availability within the robot-operator's hours of operation. Participants could also initiate calls to LDC's robot-operator during its hours of operation. For each call, the robot-operator collected information such as the time of the call and ANI. Participants who initiated calls entered PINs. The robot-operator attempted to prevent any pair of subjects from speaking more than once. However, given the small number of participants and the large number of calls made some repeat pairings were inevitable.

Once approximately 175 subjects were recruited, they were marked as active and the robot-operator was opened up for calls. A mass email was sent out to all participants informing them that the study had started and reminding them of the availability times they had specified when the system would call out to them. If these times were no longer possible, subjects were asked to contact LDC and arrange for a different schedule.

The robot operator was activated on October 7, 2008. During the project the LDC staff's primary tasks were to answer phone calls from participants, help them work through any problems they were experiencing, take part in calls from the office when necessary and audit the phone calls coming in each day for speaker ID, signal and conversation quality.

Of the 209 registered participants, 180 (86%) made calls. Within six weeks the study had reached and surpassed the goal of 100 subjects with at least 10 calls and 25 with 20 calls. The number of calls made and audited was reviewed daily and when it appeared after 4 weeks that attaining the goal was imminent, a mass email was sent informing active participants that we would be shutting down the platform. They were thanked for their

participation and informed that they had two weeks in which to finish making their calls. The two-week time frame allowed participants who had not yet started making calls to finish. Those who had made any effort would be able to comfortably. Two weeks later the collection was complete and at midnight on November 17, 2008 the robot operator was shut down.

### Call Topics

Before subjects agreed to a call, the robot operator provided a brief description of the topic of the day. The purpose of the topic is to break the ice between subjects who do not know each other and to help vary the vocabulary used from call to call. Without such prompts the principal topic of discussion becomes the study itself. Subjects could decline the call based on the subject, but they were also informed that they could discuss any topic on which they agreed and that there was no penalty for conversations that strayed from the assigned topic as long as both subjects could agree on a topic and converse reasonable. Once the subject-pair had been connected, the robot operator described the topic of the day and began recording. All new topics and descriptions were developed and recorded at LDC for this study. With a total of 66 different topics and less than 40 days of calling it was unlikely that participants would be asked to speak twice on any one topic.

### Call Duration

Each call had a 10-minute duration. Subjects were informed of this, given a warning that their time was running out and given a chance to record a comment after the call was completed. The robot operator stopped recording and disconnected the participants at the end of ten minutes.

### Call Logging

The robot operator logs information about each attempted call, both outbound and inbound, and about all successful pairings. Subject PINs and the time of call were saved in a database that also includes a pointer to the speech file containing the audio of the relevant side of the call.

## 4. Auditing

Call auditing progressed in parallel with call collection. Auditing involved listening to parts of each call to assure that the speaker associated with each PIN is consistent and to indicate the levels of background noise, distortion and echo present. Each auditor was presented with the entire side of a call on which to base decisions. They could quickly establish callers' identity by comparing the voice to previous calls using the same PIN or ID. The tool used by the auditors allowed them to scan a wave form for the whole side, which made it relatively easy to establish the amount of speech, poor quality, or technical interference. Calls that fell short of the minimum number of minutes of speech (6) were automatically dismissed by the system and did not appear in the audit queue. Staff could track the number of calls attempted, both those placed by the participant and those placed to the participant by the system. Each subject record was set with a maximum number of calls, in this case 12, that would be allowed to pass audit. Once the participant had reached their goal the system would no longer call out and they were not able to initiate calls to other participants. At any time during the calling cycle, participants could call LDC staff or send and email to query their status.

## 5. Delivery

### Accumulating Historical Call Data

Preparation of our collection for delivery and publication involved collating each participant's records with records of any prior activity. Many of our subjects had more than one subject ID from different projects. For these subjects we combed through their identities, selected one to keep as the reference, and subsumed all the other IDs under the reference ID. Staff then deleted the legacy subject IDs. Once this step was completed subjects' calls from previous studies were linked with current calls under their unique subject ID.

Prior to release of the data to the sponsors, a second level of auditing was conducted in which all calls for each participant were made available. Calls from prior studies as well as those recently collected were listed. The auditor could then listen to a short portion of each call to verify that in fact all calls attributed to a subject were from the same person.

### Yield

The final data delivered to NIST for use in SRE2010, and presumably the same to be published once its use in common task evaluations is complete, has the following characteristics.

- 171 subjects are included
- gender split is 58% female, 42% male
- 78% have some education beyond high school
- 97% are native speakers of English
- 71% are Caucasian, 13% African American, 5% Latino, 5% Asian, 1% Middle-Eastern, 5% no answer
- 28% were smokers

Figure 1 shows the Greybeard yield in terms of the count of subjects on the y-axis and the number of calls they completed on the x-axis. As can be seen, the Greybeard project exceeded its target of at least 100 subjects completing at least 10 calls and at least 25 subjects completing 20 calls.

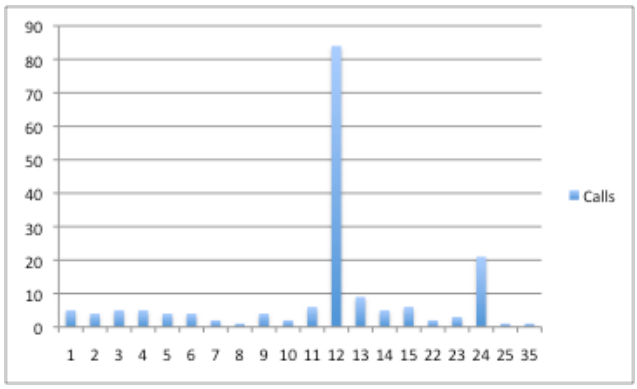


**Figure 1: Greybeard Subjects by Calls Completed**

## 6. Conclusion

The data resulting from this collection should serve as an ample resource for researchers wanting to explore the effects of aging on voice. A subset of this data was in use in the NIST 2010 Speaker Recognition Evaluation (SRE) at the time this paper was written and the results of this first use will be public by the time this paper is presented. Like all LDC data, it will be published in the LDC Catalog and shared generally once it has been used in the relevant open NIST SRE campaigns. It is our hope that these data will encourage speaker recognition researchers to explore the effects of aging on voice.

## 7. Acknowledgements

This effort was sponsored in part by the Central Intelligence Agency.

## 8. References

Biever, Dawn, Dianne Bless, (1989) Vibratory Characteristics of the Vocal Folds in Young Adult and Geriatric Women, Journal of Voice, Volume 3, Issue 2, June 1989, Pages 120-131.

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, Crosschannel Speaker Recognition Research, LREC 2006: Fifth International Conference on Language Resources and Evaluation.

Cieri Christopher, Linda Corson, David Graff, Kevin Walker (2007) Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora, Interspeech 2007, Antwerp, August 2007.

Decoster, Wivine, Frans Debruyneartin (1990) Longitudinal Voice Changes: Facts and Interpretation, Journal of Voice, Vol.14, No. 2, pp. 184-193.

Godfrey, J. J., Holliman, E. C. & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, volume 1, pp. 517–520, San Francisco.

Miller, David, Christopher Cieri, Kevin Walker (2001), Switchboard Cellular Resources for Speaker Recognition, NIST Speaker Recognition Workshop, Maritime Institute of Technology and Graduate Studies, Linthicum MD, March 2001.

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

Vipperla, Ravichander, Steve Renals, Joe Frankel (2008) Longitudinal Study of ASR Performance on Ageing Voices, ISCAA, Proceedings of Interspeech, September 22 - 26, Brisbane Australia.

Yuan, Jiahong, and Mark Liberman, (2008) Speaker Identification on the SCOTUS Corpus, Proceedings of ASA 2008.