Speaker Recognition: Building the Mixer 4 and 5 Corpora

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

Abstract

The original Mixer corpus was designed to satisfy developing commercial and forensic needs. The resulting Mixer corpora, Phases 1 through 5, have evolved to support and increasing variety of research tasks, including multilingual and cross-channel recognition. The Mixer Phases 4 and 5 corpora feature a wider variety of channels and greater variation in the situations under which the speech is recorded. This paper focuses on the plans, progress and results of Mixer 4 and 5.

1. Introduction

The Mixer style platform is designed to respond to new issues in telephone data collection. Prior to the development of Mixer, it was possible to recruit subjects and offer a free long distance call as compensation. Most subjects had one land-line in their home, another at the office and cell-phones were a rarity. Within the last 10 to 15 years this has changed drastically. Cell-phones are ubiquitous, and the plethora of calling plans that make long distance, even international calling quite inexpensive made our previous model out-dated. What has also changed is the advent of call-screening and call forwarding. It was simply more and more difficult to reach someone at home by phone. While prior models had employed a robot operator that would call participants at days and times they designated, the Mixer model allowed participants to dial in at their convenience as well and be connected to another participant automatically. The use of this model increased the efficiency of the calling platform and decreased the cost per call. In addition to continuing the practice of recruiting many more subjects than required, Mixer also initiated the practice of setting subjects' goals 20-25% higher than that required by project sponsors. Mixer also adjusted subject compensation from a flat rate per call to a smaller per call payment, a minimum number of calls necessary to receive any payment as well as a bonus for completion of all targets.

In order to support a variety of new research tasks the Mixer studies were designed initially to include multi-lingual and cross-channel collection. In the former bilingual or multilingual speakers are asked to make some calls in English and some calls in one of the non-English languages selected for that study. In the latter subjects complete calls from a special location where the impulse response has been measured and that is equipped with a multi-channel recording system.

Mixer studies include a number of tasks. Most require the core collection of a small number of short calls from a large number of subjects. To support the collection of calls from unique handsets conditions, subjects are asked to make four calls from handsets they use exactly once in the study. Once a handset reappears in the study, it is no longer considered unique. Extended Data refers to collection of 20 or more calls per subject.

2. Multi-Channel Set-up

The microphone configuration built for use in Mixer 4 and Mixer 5 includes a multi-channel digital interface, notebook computer and multi-channel pre-amp with the capacity to handle 16 channels though only 14 were used in the recordings. Table 1 summarizes the kinds of microphones used and their placement. One is devoted to the interviewer. The remaining microphones are devoted to the subject and are used and placed consistently in each interview session across locations.

Ch	Microphone	Placement	Subject /Refrence
01	Shure MX185 Lavalier	Interviewer	
02	Shure MX185 Lavalier	Subject	
03	Etymotic Micro-array	Interviewer	
04	Shure MX418X Podium	Desk Front	Center
05	Crown PZM-6D	Desk Top	Center
06	Audio Technica AT3035	Desk Front	Right
07	Audio Technica Pro45	Hanging	Center
08	Panasonic Camcorder	Desk Top	Right
09	RODE NT6	Desk Front	Far Left
10	RODE NT6	Desk Front	Center Left
11	RODE NT6	Desk Front	Center Right
12	RODE NT6	Desk Front	Center Far Right
13	AcoustiMagic Array	Wall Mounted	Center
14	Lightspeed Headset	Subject	

Table1.Microphones for Cross-Channel Recordings; Mixer 4 Cross-Channel Calls, Mixer 5 Interviews

3. Mixer Call Platform

For Mixer 4 & 5, all calls, including the cross-channel calls utilized the Mixer calling platform. Subjects would either receive a call or they could dial in using a toll-free number and be connected with another participant from Mixer 4 or 5. Since both studies ran simultaneously and neither had a large participant pool, when a reasonable number from both projects were registered the platform was opened. In order to avoid repeated pairings the platform was not opened until the pool of participants had reached 200. The robot operator announced the topic choice of the day taken from a list of 40 topics. Topics were cycled so that it was unlikely for participants to asked to discuss a topic they had previously encountered. Participants generally attempted to discuss the topic provided, but there was no penalty for conversation that was not on topic as long as it was usable dialogue. There was a range of topics available dealing with timely and interesting issues that were designed to elicit speech. Among them were political, cultural, social and religious topics. Participants could also be asked to discuss such things as hobbies, and likes and dislikes in music. Participants could also refuse to take the call after hearing the topic of the day. Participants were limited to one call a day. The system would not connect them until the appropriate time had passed. All calls received were audited for length, quality of sound and quantity and suitability of speech. In Mixer 5 it was possible for non-English speakers to connect and they were encouraged to converse in their native language. These calls were audited by a native speaker of that language. Once the calls were audited they were accepted into the database of calls and participants could log into their account and track their progress. Once participants reach their goals, their accounts are marked as inactive and are no longer able to dial-in and the system does not automatically call-out.

4. Design and Goals for Mixer 4

The purpose of Mixer 4 is to support speaker recognition research and technology evaluations. Mixer 4 consists of core telephone and cross-channel data. All subjects are required to be native speakers of American English. Specifically, 400 subjects made 10, 10 minute calls. Of these 400, 200 visited one of two sites where they made 2 telephone calls while also being recorded on a cross-channel platform. Consequently, at least 25% of the participants were from the Philadelphia area, 25 % were from the Berkeley area and 50% were recruited from all over the US. In an effort to balance the dialects we specifically recruited participants in Texas, Georgia, Illinois and New York. Of the 400 total, 100 participants are asked to make an additional 10 calls, for an extended-data set of 20 calls for 100 participants. Cross-channel calls were made using platforms designed by the Linguistic Data Consortium (LDC) and put in place at LDC in Philadelphia, Pennsylvania and at International Computer Science Institute (ICSI) in Berkeley California. The hardware, software and set-up were precisely duplicated at each location. Approximately half of the 200 cross-channel calls were made at each location. This increases the variety of dialect and vernacular in the recordings.

5. Design and Goals for Mixer 5

Mixer 5 focused on cross-channel recordings of face to face interviews where the goal is to elicit speech within a variety of situations. Specifically, 300 subjects each complete 10 calls and 6 interview session. See Table 2 for session breakdown by Speech Act. Interview participants include a subject, and interviewer and in the sessions that included telephone calls, a confederate. The interviewer engages the subject in conversation and guides him or her through a series of speech elicitation exercises. The confederate's role is to assist the elicitation of speech characterized by high or low vocal effort as discussed below.

Each subject participates in 6 thirty minute interview sessions spread over at least three days, preferably more, with at least 30 minutes rest between sessions that occur on the same day. The goal of these sessions is to record speech n a variety of situations that vary formality and model multiple naturally occurring performances and interactions The goal of the informal interview sessions is 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 thus increasing the probability that the subject's language approximates his or her vernacular. The more formal elicitations are intended to elicit speech that is either phonetically rich or else focused upon specific linguistic phenomena. To encourage the production of vernacular speech, the formal elicitation is deferred until the second session of six. The profile of the session is expected to be generally formal at the beginning of the first session, with formality generally decreasing into the second session. The interviewer leads the subject through the informal sessions by asking a series of questions. At the beginning of each line of questioning the interviewer watches for signs of interest on the part of the speaker, pursues topics of interest and abandons that produce no response or produce signs of uneasiness. Where appropriate the interviewer encourages the subject to tell stories about events in the subject's past and to describe objects or procedures in detail.

In order to elicit multiple repetitions of a small amount of speech in which the same words appear, each of the six sessions begins with the subject answering the same questions. In many cases the subject will have just met the interviewer for the first time, entered an unknown environment and completed paperwork. As a result he of she may be hesitant in conversation and prone to formality. Respecting this, a warm-up follows with the kind of conversation characteristic of first meetings, discussion about the subject's travel to the interview site, the weather and similar superficial topics. The next session of the interview focused on the personal and family history of the subject. The interviewer asks questions which focus on demographics such as, where the subject was born, grew up and went to school and what the subject is currently doing for a living.

Informal conversation makes up a large portion of the study and spans all of the interview sessions. The interviewer engages the subject in informal conversation exploring a variety of topics in search of those that ignite the subject's interests. In Transcript Reading the subject, using a natural speaking voice and style, reads individual utterances taken from transcripts of phone conversations collected in earlier studies at the LDC. All of the earlier utterances were audited, transcribed and then culled to produce the list. In Story Reading the subject reads stories containing phonetically balanced text. For Sentence Reading, the subject reads a subset of the TIMIT sentences in a natural reading voice and style. In Phrase/Word List Reading, the subject reads from lists that are designed to elicit speech that will produce speech in which the vernacular is most easily heard. This is generally seen in the characteristic production of vowels. In the Low Vocal Effort Call the subject participates in a brief (5 minute) telephone call characterized by low vocal effort as a natural result of a loud and clear telephone circuit in which the subject's voice is feed back through the headset.

Session Number	1	2	3	4	5	6	Min
Repeating Questions	1	1	1	1	1	1	6
Warm-up	4						4
Family Personal	5						5
Informal Conversation	20	9	14	9	9	10	71
Transcript Reading		20	15	10	15	10	70
Story Reading				5			5
Sentence Reading					5		5
Phrase/Word List Reading						5	5
Low Vocal/Effort				5			5
High Vocal/Effort						4	4
Total /Session	30	30	30	30	30	30	230

Table 2. Breakdown of Minutes/Speech Act/Session

5.1 Mixer 5 High and Low Vocal Effort Calls

Approximately one third of all Mixer 5 interview subjects are asked to make two 5 minute phone calls. These calls are scheduled as part of interview session 4 and 6. During these calls the subject wears Lightspeed XLC-20 headphones which provide 40dB passive acoustic isolation from ambient room noise. The subject's speech if reintroduced through the headphones electronically (side-tone), mixed with the remote speaker's speech and a white noise signal; the cumulative level of this mixed input audio is 65dB and the relative levels of the mix components are 30% side-tone, 40% remote speaker and 30% white noise. The white noise is generated by a software application running on the laptop. The audio output of the laptop is connected to a mixer along with the remote speaker's connection and the subject's microphone. The addition of the noise causes the participant to increase their vocal effort. For the low vocal effort the subject hears their own voice through the headphone which automatically causes them to reduce their effort.

6. Data Collection

Recruiting for Mixer 5 was done almost entirely by word of mouth and required virtually no advertising. Recruiting for Mixer 4 entailed a small amount of advertising, all of which was conducted via web-resources such as Craigslist.

Registration was conducted via the projects web-pages at the LDC and could be completed prior to the visit or once the participant was at the site. During the registration process participants were asked to give their names, addresses, gender, level of education, country where born, country where raised, native language, other languages spoken and age. All demographic information that is collected is used for payment purposes only and is not published as part of the corpora. Any personal identifying information is kept separate from the recordings. Upon completion of the registration, each participant is issued a unique Personal Identification Number (PIN). Using this number, participants may access their record at any time to monitor their progress. The PIN is their means of identification when placing or receiving phone calls.

At the LDC, active on-site collection for Mixer 5 began in February of 2007, and Mixer 4 in May 2007. At the ICSI location, Mixer 5 began In April 2007 and Mixer 4 in August 2007. The calling platform was opened as soon as the participant pool reached 200. Active collection was completed at both locations and the platform was closed on December 21, 2007. At both sites, there was a small number of staff who trained in conducting socio-linguistic interviews. There was also staffing at the LDC site for planning, project management, participant care, auditing and technical and database programming.

Calls from both projects are kept in a master database that is linked to subject data. Interview files are also linked to the the database. The common identifier for the records being a unique subject identification number that is separate from the PIN. Each file is date and time stamped, includes the subject ID, site location for interviews and multi-channel calls as well as session number for interviews.

7. Conclusion

The most difficult aspect of these collections was scheduling of participants to visit the sites. The logistics of scheduling 300 participants for multiple days of interviews and an additional 200 participants to place multi-channel calls took great planning, careful record keeping, diligence and patience. Scheduling was generally conducted by phone and follow up scheduling was conducted at the end of each session. Both the LDC and ICSI restricted on-site visits to regular Monday through Friday business hours. Fitting everyone in was a logistical coup. It is our hope to conduct additional projects using the multi-channel set-up and the interview protocol.

8. References

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

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

Christopher Cieri, Joseph P. Campbell, Hirotaka Nakasone, David Miller, Kevin Walker
The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data REC 2004: Fourth International Conference on Language Resources and Evaluation, Lisbon