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

Christopher Cieri, Linda Corson, David Graff, Kevin Walker

Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA
{ccieri|corsonl|graff|walkerk}@ldc.upenn.edu

Abstract

This paper describes new language resources designed to support research in speaker recognition. It begins with a brief overview of collections protocols, motivates the shift from the Switchboard protocol to the Mixer protocol, summarizes yields from the earliest phase of Mixer collection and then describes more recent phases, yields and expected yields and lessons learned.

Index Terms: linguistic resources, speaker recognition, multilingual, cross-channel, intrinsic variation

1. Introduction

Recent progress in speaker recognition technology, early or imminent adoption of that technology for forensic and commercial applications, and increased interest in biometrics to support current and emerging security needs have led researchers in speaker recognition to seek new challenges and to develop new approaches to meet those challenges [1,2,3]. This trend led in turn to the creation of the original Mixer corpus, Mixer Phase 1. Since that time the Mixer corpora, Phases 1, 2, and 3, have evolved to support an increasing variety of research tasks, including multilingual and cross-channel recognition. and have featured in the 2004, 2005 and 2006 NIST Speaker Recognition technology evaluations [4,5,6]. Collection of Mixer Phases 4 and 5 is currently underway. These corpora feature a wider variety of channels and greater variation in the situations under which speech is recorded. Since Mixer Phases 1 and 2 have been described adequately elsewhere [7], this paper will only briefly summarize their characteristics after an overall introduction to the Mixer franchise and will then focus on the results of Mixer 3 and plans and early progress of Mixer 4 and 5.

The Mixer collection protocols were designed to respond to several developments that emerged at the start of the new millennium. Collections of conversational telephone supporting speaker recognition research typically contain several (8 or more) relatively short calls (5-6 minutes) from a relatively large number of speakers (400) who do not know each other and who speak on assigned topics. The Switchboard protocol had previously driven such collections. Under that protocol, subjects registered times available to participate in a call and identified topics of interest from those provided by the organizers. Whenever some registered subject initiated a call to the robot operator that managed the study, the robot operator called other participants in the hope of finding a match. If the robot was able to match speakers, it played a description of a selected topic and then recorded the next 5 or 6 minutes of the conversation.

By the end of the 1990's, the Switchboard protocol had become complicated by a series of constraints imposed to try to improve resulting collections. Where previous Switchboard collections adopted the goal of a certain number of subjects who completed a certain number of calls on average, later collections required a minimum number of calls from each of the target number of subjects. Instead of requiring, say, 4000 calls from 400 subjects, later Switchboards required 10 calls from each of 400 subjects. Furthermore, the robot operator that handled calls was programmed to wait for an inbound call from a subject at which time it used a single line to try to find a conversation partner calling one available subject after another until it succeeded or the initiating caller hung up. During this search the robot operator avoided contacting subjects to whom the initiating caller had already spoken and tried to pair subjects by the topics in which they had previously expressed interest. These constraints, while imposing conditions that should be met in test sets built for the NIST evaluations, also complicated the collection. At the same time, telephone behavior was changing relative to the decade in which the Switchboard protocol was conceived. Telephone users were increasingly switching to cellular phones with voice mail and call forwarding and *screening calls* was becoming a household term. The combination of this shift in behavior and the constraints mentioned above had complicated collections and driven up the cost per successful call by the time the fifth Switchboard collection, Switchboard Cellular Phase 1 was underway. Responding to this increase in cost, Cellular Phase 2, the last collection to use that protocol dropped the constraint on matching subjects' topics of interest and extended the practice of recruiting many more subjects than were required to finish. Meanwhile, in the DARPA EARS program, the Fisher conversational telephone speech collection had turned previous practice on its head by employing a robot operator that initiated calls to all available subjects [8]. This new approach dramatically reduced the cost per successful call and the time required to collect. The Mixer protocol continued the practice of calling subjects according to an availability schedule and at telephone numbers they provided but also allowed subjects to initiate calls at the time and from the handset of their choosing. 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 followed by a large completion bonus. Finally, Mixer also removed the constraint against repeated pairings of subjects. These adjustments had the desired affect of reducing costs. To date, the Mixer collections have had the lowest cost per successful call of all LDC speaker recognition collections.

2. Mixer Goals

In order to support a variety of new research tasks the Mixer studies were designed initially to include multilingual 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 the 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 systems.

Mixer studies include a number of separate tasks. All studies require the core collection of a small number of short calls from a large number of subjects. In several Switchboards, subjects were asked to complete calls within a variety of environments including quiet offices, public places and moving vehicles. This requirement was dropped before the first Mixer collection was underway. To support the unique handset conditions, subjects are asked to make four calls from handsets that 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. In Transcript Reading each of a subset of subjects reads samples from transcripts of their calls and calls from other subjects. Table 1 summarizes the tasks completed or planned in each of the Mixer phases .

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

Table 1: Tasks within Switchboard & Mixer collection efforts

3. Mixer 3

The Mixer 3 collection was initiated to address two needs. First LDC was, at the time, engaged in a collection of conversational telephone speech to support Language Recognition. The protocol used in that case was a variant of the CallFriend protocol in which subjects completed a single call to a friend or family member within the continental United States or Canada on the topics of their choosing. The call was toll-free and both caller and callee were compensated. This protocol had worked well through the 1990's when it was used to collect more than 1000 calls in more than a dozen linguistic varieties including: American English, Canadian French, Egyptian Arabic, Farsi, German, Hindi, Japanese, Korean, Mandarin, Russian, Spanish, Tamil and Vietnamese. However, the new collection was running more slowly than desirable. The causes were presumably the lack of incentives. The free phone call was not worth as much as it had been and the compensation at 1 USD per minute was not incentive enough for participants since calls were limited to only 10 minutes in length. At the same time, there was a need for new data to support the NIST Speaker Recognition evaluation of 2006. It was expected that a Mixer collection could meet both needs because it had been observed previously that there is a bimodal distribution of speakers with respect to the number of calls completed. Many subjects make 0 calls or 1 call and drop out of the study. Of the remainder approximately 70% accomplish 80% of the established goals. With careful recruiting of speakers of the target languages it was assumed that those who made 1 call before dropping out would still provide useful data for the Language Recognition evaluation while those who complete the target number of calls would provide a single call for language recognition and the remainder of their calls for speaker recognition. In order to maintain robust evaluations, this would require that the calls used for the first evaluation not be released until the second evaluation were complete. Mixer 3 performed as expected. Where the previous CallFriend protocol generated a small number of calls most of which were useful for language recognition, Mixer generated a large number of calls most of which were useful for speaker recognition with a smaller percentage useful for language recognition. Specifically, more than 2900 Mixer 3 subjects each made a call in one of 19 languages including Bengali, 4 dialects of Chinese, 3 dialects of English, Farsi, Hindi, Italian, Japanese, Korean, Russian, Spanish, Tagalog, Thai, Urdu, and Vietnamese. For speaker recognition, 3918 subjects completed 19,951 calls. Of these, 1867 subjects have completed 15 or more calls. Because there was a seamless transition from Mixer 3 to Mixer 4 and 5, some of the 1867 subjects who can complete Mixer 4 and 5 tasks will be consider part of those studies. At LDC, we continue to use this "piggybacking" approach to collect low cost language recognition data along with speaker recognition data.

4. Mixer 4 Cross Channel Calls

To support speaker recognition research and upcoming technology evaluations, Mixer 4 will focus on cross channel data. Specifically 400 subjects will make 10 short phone calls; 200 of those will visit one of two sites where they will complete 4 telephone calls while also being recorded on the cross-channel platform. The 8 microphone configuration built for Mixer 1 and 2 [8] has been replaced with a system that includes a multi-channel digital interface, notebook computer and multi-channel preamp with the capacity to handle 16 channels though only 14 are in use currently. Table 2 summarizes the kinds of microphone connected to the 14 channels and how they are placed. One is devoted to the interviewer. The remaining microphones are devoted to the subject and are used and placed consistently in each interview session. Work on Mixer 4 is already well underway. The cross channel recording platforms are built and deployed at LDC and ICSI; the former has been thoroughly tested and the latter is undergoing testing now. Subject recruitment is underway. Once there are 200 subjects in the pool, call collection will begin in earnest,

#	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 micro-array	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 Center Left
11	R0DE NT6	Fixed: Desk Front, Subject's Center Right
12	R0DE NT6	Fixed: Desk Front, Subject's Center Far Right
13	AcoustiMagic Array	Fixed: Wall Mounted, Subject's Center
14	Lightspeed XLC-20	Worn: Head Mounted, Only During Calls

Table 2: microphones in Mixer 4 & 5 cross-channel collections

5. Mixer 5 Interviews

To support yet another new challenge, Mixer 5 will focus on cross-channel recordings of face to face interviews where the goal is to elicit speech within a variety of situations. Specifically 300 subjects will each complete 10 calls and 6 interviews sessions. Interview participants include a subject, an interviewer and a confederate. The interviewer engages the subject in conversation and guides her or him through a series of speech elicitation exercises. The confederate's role is to assist in the elicitation of speech characterized by high and low vocal effort speech discussed below.

Each subject participates in 6 thirty-minute interview sessions spread over at least 3 days with at least 30 minutes rest between any two sessions occurring on the same day. The goal of these sessions is to record speech in a variety of situations that vary formality and model multiple naturally occurring performances and interactions. Structurally the sessions consist of a series of informal interviews punctuated by more formal elicitations. The goal of the informal interviews 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 approximated 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 series of questions. At the beginning of each line of questioning, the interviewer watched for signs of interest on the part of the speaker, pursues topics of interest and abandons topics 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.

The structure of the interview sessions follows below:

Session 1
- Repeating Questions
- Warm-Up
- Family and Personal History
- Informal Conversation

Session 2
- Repeating Questions
- Informal Conversation
- Transcript Reading

Session 3
- Repeating Questions
- Informal Conversation
- Transcript Reading

Session 4
- Repeating Questions
- Informal Conversation
- Transcript Reading
- Story Reading
- Low Vocal Effort Phone Call

Session 5
- Repeating Questions
- Informal Conversation
- Transcript Reading
- Sentence Reading

Session 6
- Repeating Questions
- High Vocal Effort Speech
- Transcript Reading
- Phrase/Word List Reading
- Informal Conversation

One goal of Mixer 5 is to elicit multiple repetitions of a small amount of speech in which the same words appear. To accomplish this, 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 or 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 section 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 currently does 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 from transcripts of previous phone conversations. In Story Reading the subject reads stories, containing phonetically balanced text, The North Wind and the Sun and Arthur the Rat. In the Low Vocal Effort Phone Call the subject participates in a brief telephone call characterized by low vocal effort as a result of a loud and clear telephone circuit being subject's voice is. In Sentence Reading, the subject reads a subset of the TIMIT sentences in a natural, reading voice and style. In the High Vocal Effort Speech the subject participates in a brief telephone call where the subject's side tone and the remote caller's voice are weak and noisy. In Phrase/Word List Reading the subject reads phrases and word lists.

To date 70 subjects have completed some portion of the interviews. Recruiting and interviewing continues with the end goal of 300 subject completing 6 sessions and 10 telephone calls. Call collection will begin in earnest when there are 200 or more subject in the pool.

6. Conclusions

The Mixer corpora have supported speaker recognition research including the NIST technology evaluations for the past three years and will continue to do so through at least 2008. The corpora feature data collected under a variety of conditions including multilingual and cross-channel collection. The first of these corpora, Mixer 1, 2 and part of 3 have been released to the speaker recognition research community. Mixer 1 & 2, having been completely exposed will be released generally starting in 2007

7. References

[1] SuperSID (2002) "SuperSID: Exploiting High-Level Information for High-Performance Speaker Recognition" SuperSID Project Final Report, Johns Hopkins University, Center for Language and Speech Processing, Reynolds, Douglas, Walter Andrews, Joseph Campbell, Jiří Navrátil, Barbara Peskin, Andre Adami, Qin Jin, David Klusáček, Joy Abramson, Radu Mihaescu, John Godfrey, Douglas Jones, Bing Xiang.

[2] Campbell, William M., Douglas A. Reynolds, Joseph P. Campbell, (2004): "Fusing discriminative and generative methods for speaker recognition: experiments on switchboard and NFI/TNO field data", in Javier Ortega-García, et. al., Odyssey 2004: The Speaker and Language Recognition Workshop, Toledo, Spain, May 31 - June 3, 2004, ISCA Archive, http://www.isca-speech.org/archive/odyssey_04, pp. 41-44.

[3] Rose, Phil (2004) "Technical forensic speaker identification from a Bayesian linguist's perspective," In Javier Ortega-García, et. al., Odyssey 2004: The Speaker and Language Recognition Workshop, Toledo, Spain, May 31 - June 3, 2004, ISCA Archive, http://www.isca-speech.org/archive/odyssey_04, pp. 3-10.

[4] NIST (2004), The NIST Year 2004 Speaker Recognition Evaluation Plan http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf.

[5] NIST (2005) The NIST Year 2005 Speaker Recognition Evaluation Plan http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf.

[6] NIST (2006) National Institute of Standards and Technologies, Speaker Recognition Benchmark Tests Page, http://www.nist.gov/speech/tests/spk/index.htm.

[7] Christopher Cieri, 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

[8] Cieri, Christopher, David Miller, Kevin Walker, (2004) "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text", in LREC 2004, Proceedings of the Language Resources and Evaluation Conference, May-June 2004, Lisbon, Portugal.

[9] LDC (2006) Linguistic Data Consortium Home Page, http://www.ldc.upenn.edu/.