The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research*

Christopher Cieri[1], Walt Andrews[2], Joseph P. Campbell[3], George Doddington[4], Jack Godfrey[2], Shudong Huang[1], Mark Liberman[1], Alvin Martin[4], Hirotaka Nakasone[5], Mark Przybocki[4], Kevin Walker[1]

1. Linguistic Data Consortium, 3600 Market Street, Philadelphia, PA 19104
2. U. S. Department of Defense, MD, USA
3. MIT Lincoln Laboratory, Lexington, MA, USA
4. National Institute of Standards and Technology, Gaithersburg, MD, USA
5. Federal Bureau of Investigation, Quantico, VA, USA

ccieri@ldc.upenn.edu, waltandrews@ieee.org, j.campbell@ieee.org, george.doddington@nist.gov, godfrey@afterlife.ncsc.mil, shudong@ldc.upenn.edu, myl@ldc.upenn.edu, alvin.martin@nist.gov, hnakasone@fbiacademy.edu, mark.przybocki@nist.gov, walkerk@ldc.upenn.edu

Abstract

This paper describes the planning and creation of the Mixer and Transcript Reading corpora, their properties and yields, and reports on the lessons learned during their development.

1. Introduction

Recent speaker identification (SuperSID 2002) research has made significant progress in meeting classic challenges, has created interest in new problems and has increased focus on forensic scenarios (Campbell et. al. 2004, Rose 2004). The NIST 2004, 2005 and 2006 speaker recognition evaluations (NIST 2004, 2005, 2006) have added crosslanguage and crosschannel tasks. Improvements in accuracy and adaptability to new languages and channels promise increased utility in forensic applications. Progress had been hampered by a dearth of appropriate data, but the situation has now improved with the creation of the *Mixer* and *Transcript Reading* corpora. This paper describes their creation and properties and reports on the lessons learned during their development.

To support research, development and evaluation of robust speaker recognition technologies, the Linguistic Data Consortium (LDC), in consultation with Lincoln Laboratory, the National Institute for Standards and Technology (NIST) and the Speaker Identification (SID) research community, created the Mixer and Transcript Reading corpora. Sponsorship and needs assessment were provided by the United States Federal Bureau of Investigation (FBI), Department of Defense (DOD) and Intelligence Technology Innovation Center (ITIC). Mixer is the label used to identify the telephone conversation collection project to its subjects as well as the corpora it yields. Within the Mixer collection project many speakers each participate in up to 30 calls of at least 6 minutes duration using unique handsets and multichannel recording devices for a subset of calls. Bilingual speakers complete at least four calls in languages other than English as well as additional calls in English. In the Transcript Reading corpus, many subjects read partial transcripts of their own and each others' previous Mixer calls.

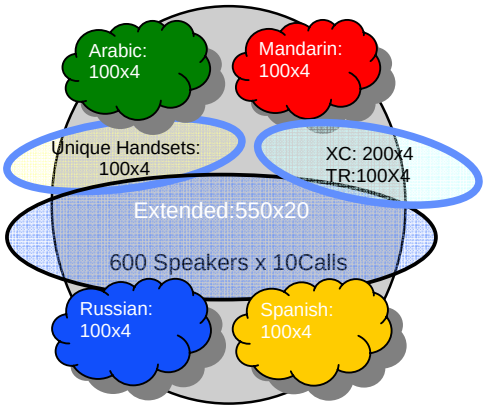

Figure 1: A summary of Mixer Phase 1 and 2 goals

To date, there have been two complete phases of Mixer collection and a third phase is underway. Phases 1 and 2 overlapped in time and subjects who had not met goals in Phase 1 were allowed to do so in Phase 2. Figure 1 summarizes the Mixer Phase 1 and 2 targets. The base goal was to record 600 subjects completing 10 calls and 550 completing at least 20. 100 subjects were to complete at least 4 calls in each of Arabic, Mandarin, Russian and Spanish and 100 were required to complete at least 4 calls using unique handsets. Finally, 200 subjects were required to complete at least four crosschannel calls and 100 of those were required to read excerpts of transcripts of their

* This work was 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.

previous calls. Given the complexity of the study, any call was allowed to satisfy more than one condition. In other words, categories were allowed but not required to overlap.

2. Methods

Mixer employed a variant of the Fisher telephone collection protocol (Cieri, et. al. 2004) in which a robot operator initiates calls to registered subjects at times and telephone numbers they specify and accepts calls initiated by subjects. The protocol connects any two available subjects fitting the constraints of the particular study.

Multichannel recording devices installed at three locations allowed subjects to initiate calls that were simultaneously recorded via eight different microphones selected and placed to represent a variety of microphone and channel conditions. The microphones were attached to a multichannel recording device (MRD). Table 1 lists the microphones used including their type or typical application, whether the microphone is balanced and shielded, the impedance and the type and source of power whether internal or external, phantom or bias.

Application	Bal,	Shield.	Imped,	Mic Pwr/Src
Studio	✓	✓	L	Phan/MRD
Podium	✓	✓	L	Phan/MRD
Hanging	✓	✓	L	Phan/MRD
Dictaphone	Line	✓	Line	Bias/Dicta.
Earboom			H	Bias/Ext
Earbud			H	Bias/Ext
PZM			H	Bias/Int Batt
Computer			H	Bias/Int Batt

Table 1: Crosschannel Microphone Types

Integrating eight varied sensors and maintaining the multichannel recording device proved more difficult than anticipated. Several sensors had to be modified and general wear on the system proved too intense for some components, which either broke or else performed below expectations.

Subjects were recruited from previous studies and via the Internet and newspapers focused on specific language communities. To compensate for expected shortfalls in participation, LDC registered 4818 subjects, all residents of North America, and set performance goals 20-25% higher than needed. Candidates registered via the Internet or telephone, provided demographic information and their hours of availability and indicated the types and numbers of all phones at which they would receive calls. Following the guidelines of the Institutional Review Board (IRB) of the University of Pennsylvania, LDC's host organization, identifying information was confidential and used for payment purposes only.

Mixer subjects were asked to participate in 12 calls speaking to other participants about assigned topics. Those who met study goals promptly were invited to continue in up to 25 calls. Subjects were given incentives to make many calls, use unique telephone handsets and speak in Arabic, Mandarin, Russian or Spanish. Subjects living near the multichannel collection facilities were invited to complete four or more crosschannel calls. Subjects who met their crosschannel goals with alacrity were invited to participate in Transcript Reading.

During the study, the robot telephone operator was active daily from 2:00PM until 12:00 midnight Eastern Standard Time, calling available subjects and receiving inbound calls. Information was collected about the time of each call and, where possible, the identifying code of the handset. Participants identified themselves via a unique number. In contrast to previous studies such as Switchboard, the protocol used in Mixer does not attempt to prevent repeat pairings of subjects, which did occur occasionally.

Before subjects agreed to talk, the platform briefly described the topic, which changed from day to day. Once two subjects were connected, the robot operator gave a more detailed description of the topic and began recording. Topics were selected from among those most successful in previous studies. Although subjects were encouraged to discuss the topic, there was no penalty for straying.

The need to match speakers of the same language, in a study where they represented less than 10% of the subject pool, required modification to the protocol. First, the logic of the robot telephone operator was changed so that it initiated outbound calls to all available speakers of a single Mixer language before calling speakers of other languages. Subjects negotiated the language of the call. All subjects were required to be fluent in English, which served as the default and the language of robot operator prompts. In addition, the robot operator was dedicated on some days to collecting calls in a single non-English language, providing a means to dynamically balance the language mixtures to meet collection goals.

Soon after collection, calls were audited to assure that the speakers were accurately identified and to log the language of the call and indicate the levels of background noise, distortion and echo present.

In the Transcript Reading corpus, 100 Mixer subjects read the transcripts of 30-second segments from their own and each others' previous Mixer calls. These readings were recorded by both the robot telephone operator and the multichannel recording device. The segments were selected to maximize the density of speech from the target subject and the lexical type/token ratio. The recordings spanned two or more sessions, each beginning with subjects reading their own transcripts. The transcripts were divided into breath groups and were displayed to subjects along with a transcript of the interlocutor's speech, which was not read by the subject. A human operator sat with the subject to catch reading errors and control the recording system. Establishing time alignment between the robot operator and multichannel recorder required additional procedures and quality control of the recordings.

3. Outcomes

The primary goal of conversational telephone speech collections designed to support speaker recognition is to record subjects completing multiple calls. Figure 2 shows, on the vertical axis, the number of subjects who completed the number of calls on the horizontal axis. The

chart shows the detailed distribution of callers by calls made in the range of 1-9 but clusters together subjects who met the goals of 10+, 20+ and 30+ calls.

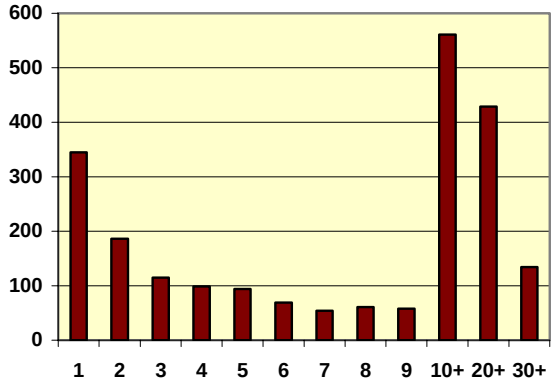

Figure 2: Callers by the number of calls completed.

Perhaps the most distinctive feature of Mixer in comparison to other corpora supporting speaker recognition research is the presence of bilingual subjects speaking in two or more. Figure 3 summarizes Mixer calls in Phases 1 and 2 by the predominant language used. Subjects spoke in the default language, English, in 84% of all calls. The other Mixer languages each account for 3-5% of all calls.

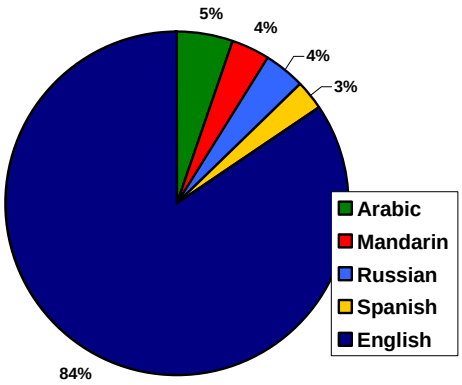

Figure 3: Calls by Language

More important than the total number of calls in a Mixer language is the number of subjects who completed a minimum number of calls, here four, in that language. Figure 4 shows Mixer subjects by the number of calls completed in Arabic, Mandarin, Russian or Spanish.

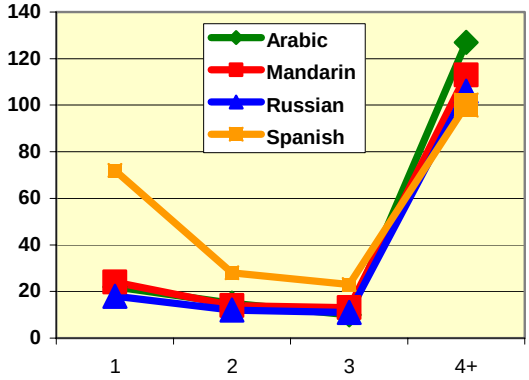

Figure 4: Subjects by number of non-English calls completed.

The compensation offered to subjects for calls using unique handsets were surprisingly effective. Although the initial requirement was to collect at least 4 unique handsets calls from at least 200 speakers, Mixer Phases 1 and 2 contain, as Figure 5 shows, many more subjects than required who completed 4 or more unique handset calls.

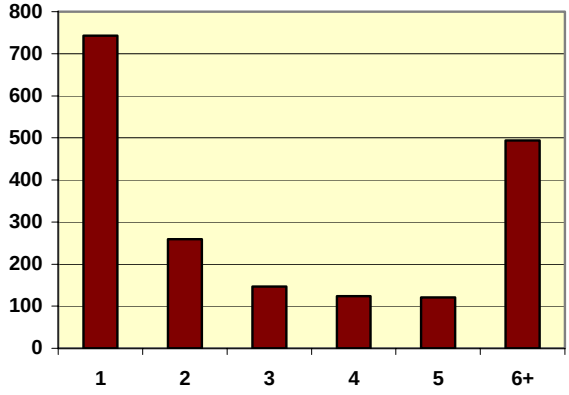

Figure 5: Callers by the number of unique handsets used.

Overall yields from Mixer were higher than expected, in some cases considerably so. Table 2 shows Mixer Phase 1 and 2 targets and actual yields. In addition to meeting or exceeding the project goals, Mixer also created a new category of subject, those who completed 30 or more calls to support speaker recognition research where extended data is available for some subjects.

	Targeted	Achieved
Base (x10 Calls)	650	1124
Arabic (x4 Calls)	100	127
Mandarin (x4 Calls)	100	113
Russian (x4 Calls)	100	106
Spanish (x4 Calls)	100	100
Extended (x20 Calls)	550	563
Super-Extended (x30 Calls)	0	134
Unique Handset (x4 Calls)	100	739
Cross Channel (x4 Calls)	200	201
Transcript Reading	100	100

Table 2: Mixer Phase 1 and 2 goals compared with yields

To reach the goals of the Transcript Reading portion of the collection, 119 subjects who had completed 4 or more crosschannel calls were invited to read 30 second excerpts from the transcripts of each others' previous conversations. 100 subjects completed this exercise

Mixer corpora contain the echo-cancelled audio of all good calls along with metadata indicating the conditions of the calls, the general demographics of the speakers, their telephone and handset types and the auditors' judgments of the sound quality of the calls. Mixer has been used in NIST's 2004 and 2005 speaker recognition evaluations and will be used again in 2006. It will then be distributed for general use.

Mixer Phase 3 is underway at the time of writing. To date 1297 subjects have completed 10,856 call sides. 415 subjects have already completed 15 or more calls. About 80% of the calls sides collected have been audited. The languages of the calls include multiple dialects of English and Chinese plus Farsi, Hindi, Italian, Japanese, Korean, Punjabi, Russian, Thai, Urdu and Vietnamese. A subset of the Mixer 3 calls collected will be used in NIST's 2006 Speaker Recognition technology evaluation. The data will also be used to support future NIST Language Identification evaluations.

4. Conclusion

The Mixer protocol has proven an effective tool in the collection of multilingual and crosschannel conversational telephone speech to support speaker recognition. Future uses of Mixer data include language identification research. Once the Mixer corpora have been fully exposed in NIST evaluations, they will be release for general use.

5. References

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.

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.

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

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

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

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

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.

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.