Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **MIT Lincoln Laboratory, Lexington, MA, USA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**2**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

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 and 2005 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, NIST and the SID research community, created the Mixer and Transcript Reading corpora. Sponsorship and needs assessment was provided by the United States Federal Bureau of Investigation (FBI), Department of Defense (DOD) and Intelligence Technology Innovation Center (ITIC). Mixer is a collection of telephone conversations from more than 2200 speakers, each participating in up to 30 calls of at least 6 minutes duration using unique handsets and multichannel recording devices for a subset of calls. At least 100 bilingual speakers completed at least four calls in each of Arabic, Mandarin, Russian or Spanish, plus additional calls in English. In the Transcript Reading corpus, 100 subjects read partial transcripts of previous Mixer calls.

* This work was sponsored by the Federal Bureau of Investigation 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.

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. Integrating eight varied sensors and maintaining the multichannel recorder 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 more than 4800 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 identified the types and numbers of all phones at which they would receive calls. Identifying information was confidential and used for payment purposes only according to procedures of the University of Pennsylvania's Institutional Review Board for the treatment of human subjects.

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 multichannel calls.

During the study, the robot 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. Unlike Switchboard, the Fisher protocol does not attempt to prevent repeat pairing 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 a given 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 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, 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 operator and the multichannel recorder. 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 interlocator'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 complete Mixer corpus contains 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.

4. 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 2005 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.

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.