The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data*

Christopher Cieri[1], Joseph P. Campbell[2], Hirotaka Nakasone[3], David Miller[1], Kevin Walker[1]

[1]University of Pennsylvania, Linguistic Data Consortium, Philadelphia, PA, USA
[2]MIT Lincoln Laboratory, Lexington, MA, USA
[3]Federal Bureau of Investigation, Quantico, VA, USA
ccieri@ldc.upenn.edu, j.campbell@ieee.org, hnakasone@fbiacademy.edu,
{damiller, walkerk}@ldc.upenn.edu

Abstract

This paper describes efforts to create corpora to support and evaluate systems that perform speaker recognition where channel and language may vary. Beyond the ongoing evaluation of speaker recognition systems, these corpora are aimed at the bilingual and cross channel dimensions. We report on specific data collection efforts at the Linguistic Data Consortium and the research ongoing at the US Federal Bureau of Investigation and MIT Lincoln Laboratories. We cover the design and requirements, the collections and final properties of the corpus integrating discussions of the data preparation, research, technology development and evaluation on a grand scale.

Introduction

This paper discusses the design and implementation of a speech corpus collection to be used for speaker recognition evaluation. In particular, it described the new Mixer corpus collection undertaken by the Linguistic Data Consortium (LDC) to support the 2004, and perhaps subsequent, NIST speech recognition evaluations and to support research and technology development ongoing at MIT Lincoln Laboratories which addresses US government needs reflected in the FBI's Forensic Automatic Speaker Recognition (FASR) prototype.

Government Requirements

Most U.S. Government forensic audio laboratories use manual and automatic forensic voice analysis investigative tools to determine the likelihood of a match between a suspect's voice and criminal's voice. The prototype Forensic Automatic Speaker Recognition (FASR) system installed at some of these Government laboratories is characterized as "text-independent" and "channel-independent" using today's cutting-edge technology. These two capabilities were set forth as the minimum requirements necessary for an automatic speaker recognition system to be even applicable under forensic conditions.

In addition to the requirements of text- and channel-independence, U.S. Government laboratories are also seeking the capability to handle non-English languages spoken by criminals and terrorists. For example, in the wake of September 11 terrorist attacks in 2001, it became clear that the U.S. Government seriously needed a new type of capability to deal with criminals or terrorists who do not speak English or who have command of multiple languages. These automatic tools need to be improved to be robust against varying languages and varying channels.

In order to facilitate future research efforts to improve FASR capability, the Government identified the following tasks: (1) multilanguage and multimodal (cross-channel) corpora collection, (2) dissemination of the corpora to the relevant research sites, (3) system performance improvement with the new corpora, and (4) system performance evaluation.

Corpus Design

In order to promote the development of robust speaker recognition technologies, we have created the Mixer corpus of multilingual, cross-channel speech. This corpus adds two dimensions to the traditional Switchboard collection: language and channels. Mixer will be used in the NIST 2004 Speaker Recognition Evaluations will prove important in assessing the state of the art of speaker recognition, and will focus research this year and in future years. The components of the MMSR corpus are enumerated below. As noted, there are necessary and desired overlaps in the collections. Mixer is a collection of telephone conversations targeting 600 speakers participating in up to 25 calls of at least 6 minutes duration. Like previous speaker recognition corpora, the calls feature a multitude of speakers conversing on different topics and using a variety of handsets types. Unlike previous studies, a large subset of the subjects were bilingual in conducted their conversations in Arabic, Mandarin, Russian and Spanish as well as English. Further distinguishing Mixer from previous studies, some calls are also recorded simultaneously via a multichannel recorder using a variety of microphones.

Mixer relies upon a collection protocol developed initially for the DARPA EARS program. Under this, "Fisher", protocol a robot operator initiates 18 calls at a time – leaving 6 lines open to receives calls – and pairs any two subjects who agree to participate at the same time. Within the Mixer collection the robot operator ordered its calls to increase the probability of pairing speakers of the same native language.

In previous call collection projects of this kind about half of all recruits have failed to participate in the

* This work is sponsored in part by the Federal Bureau of Investigation under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

study and about 70% of those who did participate achieve 80% of the stated goals. To compensate for shortfalls in participation, we recruited more than 2000 subjects, set performance goals 20-25% higher than needed and further offered per-call incentives, completion bonuses and lotteries to encourage subjects to provide the different types of data required. Specifically, subjects who completed calls on unique handsets or multimodal recording devices or in foreign languages received per-call incentives. Subjects who completed target numbers of calls in these categories received completion bonuses. Each subject who completed the base collection also received a chance in the participant lottery.

Candidates registered via the Internet or phone providing demographic data and an availability schedule and describing the handsets on which they would receive calls. The personal information candidates provided to allow us to issue payment is kept confidential and not delivered with the research data.

During the collection, the LDC robot-operator functioned daily from 2:00PM until 12:00 midnight Easter Standard Time allowing for maintenance in the morning hours. At the top of every hour, the robot operator began to call every subject who has agreed to receive calls at that time using the telephone numbers the subjects registered. Once a subject completed a call they became ineligible for 18 hours. Subjects who refused a call also became ineligible for 18 hours. A subject who did not respond to a call became ineligible for one hour. Subjects also initiated calls at their discretion. Each time the robot operator identified a pair of subjects willing to speak – whether these subjects initiated or received the call – it recorded the time of the call, the identifying codes of the handsets (ANIs) and the identifying codes of the subjects (PINs). In contrast with previous speaker recognition corpora, the robot operator did not prevent a specific pair of subjects from speaking more than once. Given the size of the study, such repeat pairings are statistically are infrequent.

In order to encourage meaningful conversation among subjects who did not know each other, we developed 70 topics of current interest after considering which topics had been most successful in previous studies. Topics ranged in breadth from "Fashionably late or reasonably early" to "Felon re-emancipation". Since Mixer required bilinguals, we attempted to balance topics of domestic interests with those having international appeal. The robot operator selected one topic each day. Subjects had the ability decline calls after hearing the topic of the day. Once a pair of subjects were connected, the robot operator described the topic of the day fully and began recording. Although subjects were encouraged to discuss the topic of the day, there was no penalty for conversations that strayed from the assigned topic.

All calls were audited shortly after collection to assure that the speaker associated with each unique identification number was consistent within and across calls, to log the language of the call and to indicate the levels of background noise, distortion and echo observed.

Core Collection for Speaker Recognition

All call activity in Mixer contributed to a core collection where our goal was 10 calls from each of 600 subjects. Knowing that studies of this kind have significant attrition rates, we recruiting over 2000 subjects and offered each speaker compensation per full-length, on-topic call with a bonus for those who completed 12 calls. In order to maximize handset variability, we also offered compensation for each call made from a unique ANI with bonuses for anyone who completed 5 such.

Extended Data

To support evaluations of the affect of volume of training data on system performance, it was desirable to have a group of subjects who completed not 10 but 20 calls. Anticipating considerable shortfall, we encouraged subjects who were so inclined to complete 25-30 calls again offering compensation per call with a bonus for subjects who exceeded 25 calls. Subjects who completed 30 calls were immediately deactivated from the study and compensated for their considerable effort.

Multilingual Data

To support the development and evaluation of systems which recognize multilingual speakers regardless of the language they speak, it was desirable to have a group of subjects who made some of their calls in English and some calls in one of the other Mixer languages Arabic, Mandarin, Russian or Spanish. For each of these languages we targeted 100 subjects who would a total of 10 calls of which 4 would be non-English. Anticipating shortfall, we required bilingual subjects to complete 12 calls of which 5 were non-English.

The robot operator clustered its outbound calls by the native language of the subjects. At any one time, it called all available speakers of Arabic before Mandarin, Mandarin before Russian and so on. Since all subjects were fluent in English, English served as the default language when, for example, the platform paired an Arabic-English bilingual with a Mandarin-English bilingual. Early in the study we learned that the persistence of the robot operator coupled with the preponderance of subjects who do not speak the same non-English language allowed the core collection to race ahead of the foreign language collection. To compensate we initiated “language-only” days in which the robot operator only allowed calls among speakers of the day’s target language.

Cross Channel Data

The goal of the cross channel collection was to record one side of a series of Mixer conversations on a variety of sensors. The sensors were chosen to represent certain target settings such as the microphones used in courtrooms, interview rooms and cell phones. Participants placed calls to the Mixer robot operator while being recorded simultaneously on the cross channel recorder. We asked participants to make at least five separate cross channel recordings.

The recording system consisted of a laptop, a multichannel audio interface, two fire wire attached hard drives, a set of eight microphones/sensors, and a simple eight channel recording application. The multichannel audio interface (MOTU 896HD) connected to the laptop via fire wire and handled eight balanced microphone connections sampling each channel at 48Khz with 16bit samples. The multichannel sensors were:

- side-address studio microphone (Audio Technica™ 3035)
- gooseneck/podium microphone typical for courtroom environment (Shure™ MX418S)
- hanging microphone (Audio Technica™ Pro 45)
- PZM microphone (Crown Soundgrabber™ II)
- dictaphone (Olympus Pearlcorder™ 725S)
- computer microphone (Radio Shack™ Desktop Computer Microphone)
- cellular phone headset (Jabra™ Earboom)
- and a second cellular phone headset (Motorola™ earbud)

The two microphones designed to be connected to the headset jack of a cell phone were modified to make them compatible with the recording hardware. The stock headsets terminate in a 2.5mm miniplug with a common ground for the earpiece and the microphone. We removed the miniplug and replaced it with a 3.5mm plug which was only attached to the microphone; the earpiece was removed from the circuit. Both headsets required bias power, which was applied using a commercial-off-the-shelf battery pack.

The Mixer Collection

Mixer call collection began in October, 2003 after we had recruited approximately 200 participants. To date (March 8, 2004), the Linguistic Data Consortium has recruited 2987 participants of which 63% are female and 37% are male. Figure 1 summarizes the percentage of the subject pool reporting native language ability in each of the Mixer languages. Some recruits reported speaking English plus 2 other Mixer languages.

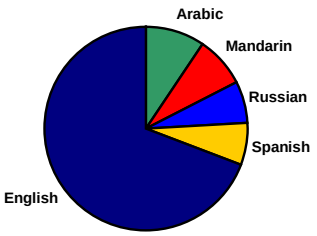

Figure 1: Linguistic ability of the Mixer recruit pool

1402 of the 2987 recruits, or 47%, have actually completed at least one full-length on-topic call. This participation rate is typical for telephone speech studies in our experience. The 1402 subjects have completed 15,254 total conversational sides (7627 calls) of which 58% contain female speakers and 42% contain male speakers. Table 2 summarizes conversations by language. most of which were collected during the "language-only" days described above.

Language	*# Conversations*
Arabic	775
English	5082
Mandarin	502
Russian	523
Spanish	744

Table 1: Mixer conversations by language

At the time of writing, we have completed our foreign language goals for all languages. We have also exceeded our extended data and unique handset goals. Specifically, we have collected 20 or more calls from 255 subjects while 247 subjects have completed 4 or more calls from unique handsets. At the time of writing the cross-channel call collection was well underway with platforms running at LDC, Mississippi State University. About one-fourth of the required 100 subjects have completed 4 cross-channel calls.

Figure 2 shows a histogram of calls by callers. The horizontal axis shows the number of calls completed while the vertical axis shows the number of callers who have complete that number of calls. The distribution one generally sees for telephone studies whose compensation effectively motivates subjects to complete the required number of calls is semi-normal with a mode just below the target number of calls and with a second mode at 1 calls. The mode at one reflect the early-rejecters common to many studies. The specific features of the Mixer study produce additional modes near the extended data goal of 20 calls and at the ceiling of 30 calls.

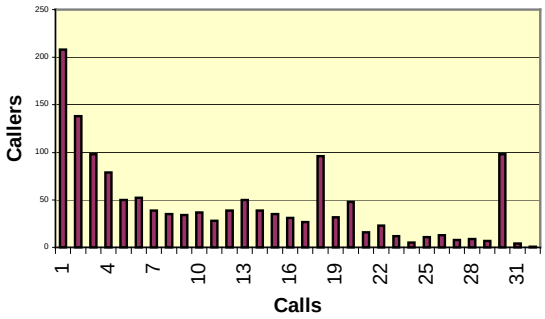

Figure 2: Histogram of Mixer callers by number of calls completed.

Because the first phase of Mixer has been very successful on a number of fronts, a second phase is now in planning. The current goals for Mixer Phase II are to double the number of cross-channel subjects bringing the total to 200 abd to accomplish a four-fold increase in the number of extended data subjects.

All of the data resulting from the Mixer collections will be used to support speaker recognition system evaluations and will be published via the Linguistic Data Consortium (http://www.ldc.upenn.edu)

Research

The MMSR corpus supports various forms of speaker recognition research and evaluation, with an emphasis on forensic-style problems [Bonastre 2003]. The MMSR corpus' three modes support research and evaluation using telephone conversations, cross microphone recordings, and transcript readings. Furthermore, the multilanguage bilingual recordings support research and evaluation on recognizing the same people talking in different languages. The MMSR corpus is the first publicly available corpus to cover all these dimensions in large-scale speaker recognition and evaluation.

The FBI's vision to sponsor the MMSR corpus to accurately reflect their speaker recognition problems and applications will focus the research community's attention on solving forensic-style problems. The MMSR corpus will support two main concerns by the FBI that have not been adequately addressed elsewhere – language and channel dependency of automatic speaker recognition.

An advanced speaker recognition research program is being conducted at MIT Lincoln Laboratory using the MMSR corpus. A robust automatic speaker recognition system will be developed to provide support to forensic analysis experts. The objective is to assist, not replace, the analyst. The goal of this speaker recognition system is to provide high accuracy across various languages and channels. MIT-LL has developed a suite of promising techniques to draw from and adapt here that have been successfully demonstrated in various evaluations. These promising techniques will be evaluated on the MMSR corpora to determine their applicability to FBI tasks. The promising techniques include high-level features, e.g., phone patterns (idiosyncratic pronunciation) and word patterns (idiosyncratic word usage) and their fusion for speaker recognition [Workshop 2002, Reynolds 2003].

Conclusions

We have described the need for robust channel and language independent speaker recognition systems and the design considerations in creating corpora to support such system development. We have also described our efforts to build the Mixer corpus that not only addresses the traditional concerns of speaker, topic and handset variation but also addresses bilingualism in multiple languages paired with English, cross-channel speaker recognition and the extended data condition. Finally, we have described the research program that the Mixer data supports.

References

Bonastre, J.-F., F. Bimbot, L.-J. Boë, J. P. Campbell, D. A. Reynolds and I. Magrin-Chagnolleau, "Person Authentication by Voice: A Need for Caution," Eurospeech, ISCA, Geneva, Switzerland, 2003.

Campbell, J. P. et al., "The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation", Proc. Odyssey 2004, The Speaker and Language Recognition Workshop, Toledo, Spain, May 31-June 3, 2004

Campbell, J. P. and Reynolds, D. A., Corpora for the Evaluation of Speaker Recognition Systems. In Proc. International Conference on Acoustics, Speech, and Signal Processing in Phoenix, Arizona, IEEE, pp. 2247-2250, 15-19 May 1999.

Cieri, Christopher, David Miller, Kevin Walker, From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields, Proceedings of EuroSpeech 2003.

Martin, Alvin, D. Miller, M. Przybocki, J. Campbell, H. Nakasone, "Data Collection for Speaker Recognition Evaluation," LREC 2004.

Martin, Alvin and Mark Przybocki, Speaker Recognition in a Multi-Speaker Environment, Proceedings of Eurospeech, 2001, Scandinavia Volume #2, Pages 787-790.

Martin, Alvin and Mark Przybocki, The NIST Speaker Recognition Evaluations: 1996-2001, Presented at Odyssey 2001.

Martin, Alvin and Mark Przybocki, Odyssey Text Independent Evaluation Data, Presented at Odyssey 2001.

Nakasone, Hirotaka, Automated Speaker Recognition in Real World Conditions: Controlling the Uncontrollable, Proceedings of EuroSpeech 2003.

Przybocki, Mark and Alvin Martin, NIST's Assessment of Text Independent Speaker Recognition Performance 2002, The Advent of Biometircs on the Internet, A COST 275 Workshop in Rome, Italy, Nov. 7-8 2002

Reynolds, D. A., W. D. Andrews, J. P. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusácek, J. S. Abramson, R. Mihaescu, J. J. Godfrey, D. A. Jones and B. Xiang, "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition," International Conference on Acoustics, Speech, and Signal Processing, IEEE, Hong Kong, 2003, pp. 784-787. Available: http://www.clsp.jhu.edu/ws2002/groups/supersid/.

"Workshop 2002 - SuperSID: Exploiting High-level Information for High-performance Speaker Recognition," 2002. Available: http://www.clsp.jhu.edu/ws2002/groups/supersid/.