The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation*

Joseph P. Campbell[1], *Hirotaka Nakasone*[2], *Christopher Cieri*[3], *David Miller*[3], *Kevin Walker*[3], *Alvin F. Martin*[4], *Mark A. Przybocki*[4]

[1]MIT Lincoln Laboratory, Lexington, MA, USA
[2]Federal Bureau of Investigation, Quantico, VA, USA
[3]University of Pennsylvania, Linguistic Data Consortium, Philadelphia, PA, USA
[4]National Institute of Standards and Technology, Gaithersburg, MD, USA
j.campbell@ieee.org, hnakasone@fbiacademy.edu,
{ccieri, damiller, walkerk}@ldc.upenn.edu,
{alvin.martin, mark.przybocki}@nist.gov

Abstract

We describe efforts to create corpora to support and evaluate systems that meet the challenge of speaker recognition in the face of both channel and language variation. In addition to addressing ongoing evaluation of speaker recognition systems, these corpora are aimed at the bilingual and crosschannel dimensions. We report on specific data collection efforts at the Linguistic Data Consortium, the 2004 speaker recognition evaluation program organized by the National Institute of Standards and Technology (NIST), and the research ongoing at the US Federal Bureau of Investigation and MIT Lincoln Laboratory. We cover the design and requirements, the collections and evaluation integrating discussions of the data preparation, research, technology development and evaluation on a grand scale.

1. Introduction

This paper discusses some of the factors that should be considered when designing a speech corpus collection to be used for speaker recognition evaluation [1]. It will specifically discuss the design of the new corpus collection undertaken by the LDC to support the 2004 and subsequent NIST speaker recognition evaluations and to support research and technology development ongoing at MIT Lincoln Laboratory which addresses US Government needs reflected in the FBI's Forensic Automatic Speaker Recognition prototype.

2. U.S. Government Needs and 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 the Federal Bureau of Investigation (FBI) is characterized as "text-independent" and "channel-independent" using today's cutting-edge GMM-UBM based ASR technology [7]. These two characteristics were set forth as the minimum requirements necessary for an automatic speaker recognition system to be applicable under forensic conditions.

However, in the wake of September 11 terrorist attacks in 2001, it became clear that the FBI seriously needs a new type of capability built into the FASR system 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.

To facilitate future research efforts to improve the FASR system's capability, the needed work includes: (1) collect multilanguage and multimodal (crosschannel) corpora, (2) disseminate corpora to relevant research sites, (3) improve system performance with the new corpora, and (4) evaluate system performance.

3. Designs

To support research and the development and evaluation of automatic systems for robust speaker recognition technologies, we have created the Mixer corpus of multilingual, crosschannel speech. Mixer is a collection of telephone conversations targeting 600 speakers participating in up to 25 calls of at least 6 minutes duration. Large subsets of the calls collected feature unique handsets and/or conversations in Arabic, Mandarin, Russian and Spanish as well as English. Some calls have also been recorded simultaneously via a multichannel recorder using a variety of microphones. Mixer relies upon a collection protocol in which a robot operator both initiates and receives calls and pairs any two subjects who agree to participate at the same time.

In previous call collection projects of this kind [2], about half of all recruits have failed to participate in the 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

* This work is sponsored 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.

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 identification 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 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 was 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.

3.1. Core Collection for Speaker Recognition

All call activity in Mixer contributed to a core collection where our goal was to collect 10 calls from each of 600 subjects. Knowing that studies of this kind have significant attrition rates, we recruited 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 calls.

3.2. Extended Data

In order to support evaluations of the affect of volume of training data on system performance, 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.

3.3. Multilingual Data

To support the development and evaluation of systems which recognize multilingual speakers regardless of the language they speak, we targeted collecting 4 calls from each of 400 bilingual subjects. Specifically, we wanted subjects who were bilingual in English plus one of Arabic, Mandarin, Russian and Spanish. For each of these languages we targeted 100 subjects who would complete 10 calls of which 4 would be 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.

3.4. 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 computer, a multichannel audio interface, two firewire-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 firewire and handled eight balanced microphone connections sampling each channel at 48 kHz with 16-bit samples. The eight multichannel sensors were:

- side-address studio microphone (Audio Technica™ AT3035)
- gooseneck/podium microphone, typical for courtroom environment (Shure™ MX418S)
- hanging microphone (Audio Technica™ Pro 45)
- PZM microphone (Crown Soundgrabber™ II)
- dictation microcassette recorder (Olympus Pearlcorder™ S725)
- computer microphone (Radio Shack™ Desktop Computer Microphone #33-3031)

- cellular phone headset (Jabra™ Earboom Radio Shack #43-1914)
- and a second cellular phone headset (Motorola™ earbud SYN8390)

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 2.5 mm miniplug with a common ground for the earpiece and the microphone. We removed the miniplug and replaced it with a 3.5 mm miniplug which was only attached to 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 crosschannel recording system is shown in Figure 1. The hanging microphone is placed high and across the room from the subject who is seated in the chair and surrounded by microphones.

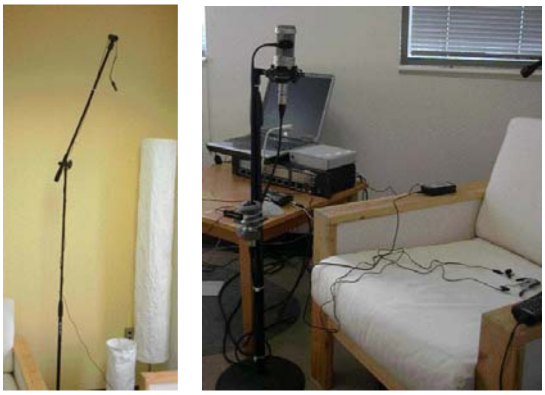

Figure 1. Crosschannel recording system.

4. The Mixer Collections

Mixer call collection began in October, 2003 after we had recruited approximately 200 participants. As of 14 January 2004, the Linguistic Data Consortium had recruited 2470 recruits of which 63% were female and 37% were male. Table 1 summarizes the linguistic ability of the subjects. Some recruits reported speaking English plus 2 other Mixer languages.

Language	*# Recruits*
Arabic	273
English	931
Mandarin	290
Russian	217
Spanish	711

Table 1: Linguistic ability of the Mixer recruit pool

1164 of the 2470 recruits have actually completed at least one full-length on-topic call. Having 47% of all recruits actually participate is typical for telephone speech studies in our experience. The 1164 subjects have completed 10,670 total conversational sides (5335 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	499
English	3437
Mandarin	463
Russian	338
Spanish	658

Table 2: Mixer conversations by language

At the time of this writing, we were approaching our foreign language goals for all languages. Figure 2 shows, for each Mixer non-English language the number of subjects who had completed 1, 2, 3, or 4 or more calls in that language. We have also exceeded our extended data and unique handset goals. Specifically, we have collected 20 or more calls from 138 subjects. 186 subjects have completed 4 or more calls from unique handsets. Currently, crosschannel call collection platforms are recording data at LDC, Mississippi State University, and the International Computer Science Institute, with plans to collect data at Rutgers University and the Georgia Institute of Technology.

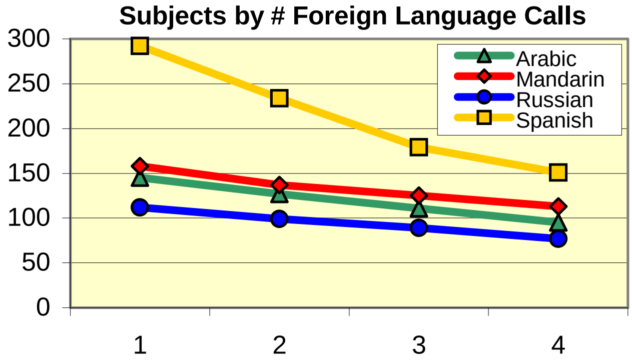

Figure 2: Mixer Subjects by the number of foreign language calls they completed.

5. Evaluations

The MMSR corpus design supports evaluation of the effect on speaker recognition performance of several factors that have received limited attention previously. These factors will be investigated in the annual NIST coordinated evaluations of speaker detection performance in 2004 and beyond [3-6, 8].

NIST plans to concentrate the evaluation on trials involving the use of different telephone handsets (at least as implied by the ANI's). This is desirable since it is hard to collect many calls involving different speakers with matching ANI's, so few same handset impostor trials could be included. Moreover, there is some increase in the amount of independent information provided by trials with different test segments involving the same speaker when the handsets used are different.

Therefore the speaker initiated conversation sides with unique ANI's will be used for test segment data, while the multiple sides with repeated ANI's will be used as sources for model training data. Note, however, that some speakers will have two or more ANI's that are repeated, perhaps because they receive calls at both home and work, or on both landline and cellular phones. Multiple speaker models may then be defined corresponding to such speakers.

5.1. Landline vs. Cellular

Many speakers made a combination of landline and cellular calls in either their training or their test calls, or both. This will support investigation of performance differences between landline and cellular data. Moreover, unlike evaluations in prior years, it will allow performance involving cellular transmission to be evaluated in the context of target trials involving the use of different telephone handsets in the training and test data. This may support progress through higher measured error rates.

In addition to comparing overall speaker detection performance on landline data with that on cellular data, it will allow comparison with the mixed conditions of cellular training and landline test segments, or vice versa. Furthermore, it will allow comparisons of the type where, for example, the target speaker set and all its model training data is fixed, while the target test segments consist solely of landline or solely of cellular data from the same speakers

5.2. English vs. Other Language

About 100 or more speakers made calls, to be used as training data, in one of the four foreign languages as well as in English. This will support investigation of whether the use of training data in a language different from that of the test data adversely affects performance. It has been generally assumed that, with traditional acoustic approaches, such language differences should have little effect on performance, but this has not been verified. The use in recent NIST evaluations of higher-level information sources, including idiolect (e.g., word bi-grams and tri-grams characteristic of specific speakers), could make within speaker language differences more problematic for achieving the best possible detection performance.

Under the collection protocol all test segment data collected will be in English. With such fixed test data, it will be possible to vary the training language for a fixed set of target speakers, in order to examine the effect on detection performance of language match or mismatch between training and test.

5.3. Telephone vs. Microphone

At least 100 speakers are being collected making calls, to be used as test data, which are recorded simultaneously over telephone lines and over several in-room microphones. This collection will be available for future evaluations.

This will support controlled comparison of the effect on detection performance of the transmission type (telephone or microphone) and the microphone type of test segment data in the context of target model training data collected over telephone lines. This stands to be of interest for applications where test data from a situation of interest might be collected in various ways, either over telephones channels or using microphones of different types placed in varying proximity to a speaker of interest, while training data is collected over a longer period of time from available conversational telephone data.

An update of the MMSR corpus, as of 26 March 2004, is summarized in Table 3 (in the style for corpora given in [1]).

Dimension	*Value*
# of speakers	690M & 972F (43%M, 57%F)
# sessions/speaker	1-36 (6 min conversations)
Intersession interval	Two within 18 hours to weeks
Types of speech	*Mixer:* conversational *Transcript Reading:* reading aloud
Microphones	Variable telephone handsets and crosschannel (multimodal)
Channels	PSTN and cellular
Acoustic environments	Home, Office, Public Space, Outdoors, Moving Vehicle
Languages	English plus bilinguals in English and {Arabic, Mandarin Chinese, Spanish, or Russian}
Evaluation procedure	Yes for NIST Evaluation sets [6]

Table 3: MMSR Corpus Description

6. Conclusions

We have described the needs for robust channel and language independent speaker recognition systems, design considerations in creating corpora to support such system development, and procedures for evaluating them. We have also described specific data collection and evaluation efforts underway at LDC and NIST, respectively. The MMSR corpus will eventually be available from LDC (please consult http://www.ldc.upenn.edu).

7. Acknowledgements

We are grateful to George R. Doddington, John J. Godfrey, and Douglas A. Reynolds for their contributions to the MMSR design.

8. References

[1] Campbell, Joseph P. and Reynolds, Douglas A., *Corpora for the Evaluation of Speaker Recognition Systems*. Proc. International Conference on Acoustics, Speech, and Signal Processing in Phoenix, Arizona, IEEE, pp. 2247-2250, 15-19 May 1999. http://www.apl.jhu.edu/Classes/Notes/Campbell/SpkrRec/.

[2] Cieri, Christopher, David Miller, Kevin Walker, *From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields*, Proceedings of Eurospeech 2003.

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

[4] Martin, Alvin and Mark Przybocki, *The NIST Speaker Recognition Evaluations: 1996-2001*, Presented at Odyssey 2001. http://www.nist.gov/speech/publications/.

[5] Martin, Alvin and Mark Przybocki, *Odyssey Text Independent Evaluation Data*, Presented at Odyssey 2001.

[6] NIST Speaker Recognition Evaluation Plans. http://www.nist.gov/speech/tests/spk/.

[7] Nakasone, Hirotaka, *Automated Speaker Recognition in Real World Conditions: Controlling the Uncontrollable*, Proceedings of Eurospeech 2003.

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