Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004

Alvin Martin[1], David Miller[2], Mark Przybocki[1], Joseph Campbell[3], Hirotaka Nakasone[4]

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

Abstract

This paper discusses some of the factors that should be considered when designing a speech corpus collection to be used for text-independent speaker recognition evaluation. The factors include telephone handset type, telephone transmission type, language, and (non-telephone) microphone type. The paper describes the design of the new corpus collection being undertaken by the Linguistic Data Consortium (LDC) to support the 2004 and subsequent NIST speech recognition evaluations. Some preliminary information on the resulting 2004 evaluation test set is offered.

1. Introduction

Evaluation of text-independent speaker recognition systems is a very data intensive undertaking. It has long been recognized in speech processing that data drives research, and that the type and quality and amount of data used to evaluate systems directly impacts the performance factors that can be examined and the statistical significance of the conclusions that can be drawn from an evaluation. Various speech corpora have been developed over the years to meet this need [1].

As conducted by NIST in recent years, each speaker recognition evaluation on conversational telephone speech has involved a corpus with hundreds of speakers, thousands of conversation sides, and tens of thousands of individual test trials. Each evaluation test set is dependent upon numerous data collection factors that affect evaluation performance. Often we wish to collect sufficient amounts of data associated with these factors so that meaningful (i.e., statistically significant) results on how these factors affect performance can be obtained. But this can lead to an explosion in the amount of data needed, so compromises are necessary.

The factors of interest, in addition to those related to the voices of the speakers themselves, include most particularly variations in the telephone handsets used and the types of transmission channels involved, and the match or mismatch of these between the training and test speech data.

Previous NIST evaluations (see [2], [3], [4], [5], [6]) have shown that performance is greatly enhanced when speakers use the same telephone handsets in their training and test data. This is not surprising since different speakers essentially always use different handsets, so success may be attained by identifying handsets rather than voices. Requiring that training and test handsets always be different is therefore a desirable evaluation objective. But collecting extensive real conversational speech data with each speaker using multiple handsets of varying types and transmission channels is a challenging endeavor.

Previous NIST evaluations have also shown how the two common handset microphone types (carbon button and electret) of landline phones affect performance. Performance is generally enhanced both by the use of electret microphones and by the use of matched type between training and test. Carbon button handsets now are becoming uncommon. Recent NIST evaluations have also shown that cellular transmission generally produces performance inferior to that with landline transmission. This is perhaps not surprising, but further investigation of related issues is needed.

2. Factors Affecting Performance

The previous NIST evaluations have made clear the need to investigate the effects of different handset and telephone transmission types on performance. The use of cellular and cordless phones has become pervasive in the past decade, and the use of specialized handsets such as speakerphones and headsets has increased. There has also been renewed interest in the effect on performance of speakers of different languages, particularly if some speakers should use multiple languages. For forensic applications there is interest on the interaction of collection channels that may include different types of microphones as well as telephone data.

2.1 Handset type

In addition to the microphone type, telephone handsets may differ in how speakers use them for speaking and listening. They may involve speakerphones, headsets, ear-buds, or just ordinary handheld devices. It is of interest to learn how these options, in different training and test combinations, may affect speaker recognition performance.

2.2 Transmission type

Landline, cellular, and cordless transmission are all widely used today. While previous evaluations have

focused on either landline or cellular calls, a careful examination of the alternatives, with training and test data always involving different handsets and sometimes involving different transmission modes, is very much needed. The effects of different types of cellular transmission are also worthy of examination.

2.3 Language

The effect of language differences on recognition performance has been a subject of great interest, but one that has received limited study, due perhaps to a lack of comparable data involving multiple languages, and especially a lack of data involving bilingual speakers.

It is generally believed that speaker recognition performance should not vary greatly with language, as os long as the speech data used is entirely in one language, but this has not been verified in a formal evaluation.[1] It is less clear what may be the effect on performance of having speech, for some speakers, in more than one language. The use of "higher level" types of features such as word n-grams, in conjunction with traditional acoustic type features, to achieve improved greater performance levels [8], as pioneered in recent NIST evaluations, could make cross-language recognition performance more problematic. But test data from bilingual speakers is needed to investigate this.

2.4 Microphones

The primary application interests for speaker recognition, especially text-independent speaker recognition, have involved voice transmission over telephone lines. This is the area of advantage that voice possesses over other biometrics. But there is some interest, particularly for forensic applications, in recognizing voices recorded over various types of microphone channels. Of particular concern is the impact on performance of training and test data being recorded over different channel types, perhaps telephone in one case and microphone in the other. This cross channel speaker recognition problem was investigated to a limited extent in the 2002 NIST evaluation using a Federal Bureau of Investigation (FBI) provided corpus (described in [9]). Further study of this matter requires more extensive cross-channel data collection.

3 Mixer Corpus

In planning for the 2004 NIST evaluation and beyond it was decided to ask the Linguistic Data Consortium (LDC) to undertake a new set of conversational telephone recordings based on the Fisher paradigm used in the past year to collect data for evaluation of conversational speech recognition in DARPA's Effective Affordable Reusable Speech-to-text (EARS) program [10], discussed in [11]. This paradigm involves an automatic platform that initiates pairings between participants who have signed up to take part in the program. They are called at phone numbers they previously specified during hours when they indicated they would likely be available to participate in short (typically six minute) conversations on assigned topics. Because of the desire to collect data with handset and transmission type variation, the paradigm was modified for the new speaker recognition oriented collection to encourage participants to initiate themselves a number of conversations using unique phone numbers. Using this "Fishboard" paradigm (combining aspects of the Fisher and the previously used Switchboard paradigms), it is hoped that 600 or more speakers will take part in ten or more such conversations, with four or more of these initiated by the individual speaker from unique phone locations. Enthusiastic subjects are to be encouraged to make 25 or more calls. The resulting corpus has been given the name of Mixer [12], [13].

A special effort has been made to recruit bilingual subjects who speak Arabic, Mandarin, Russian, or Spanish in addition to English. When someone speaking one of these other languages is called, an attempt is made to pair this speaker with another who speaks the same language. Speakers are instructed to talk in one of these four Mixer languages if they both are able to do so, and in English otherwise. Thus a significant percentage of the calls by the bilingual speakers should be in a language other than English.

Table 1 provides collection figures at the conclusion of the first phase of Mixer. While the recruit numbers are large, note that some recruits end up either not contributing or contributing only a few conversations, and thus become of limited use for evaluation of speaker recognition systems.

Language	Recruits	Conversations
Arabic	317	774
English	1120 (not bilingual)	4968 (by all speakers)
Mandarin	317	502
Russian	262	520
Spanish	878	742
Total	2894	7506

Table 1: *Mixer Corpus collection statistics at the conclusion of its first phase*

Each speaker is asked in each call to specify the phone transmission type (cellular, cordless, or regular landline) and the handset type (speakerphone, headset, ear-bud, or hand-held). This self-reported information could later prove valuable in sorting out the effects of these factors on recognition performance. Information is also being collected from each speaker on his or her place of birth, age, and level of education.

A special collection effort was initiated to collect cross-channel conversational speech data as part of the overall Mixer collection. Three sites were designated as locations where 35 people were to be recruited to each participate in five conversations. The conversations were to be made

[1] A previous NIST evaluation included a test on the Spanish language AHUMADA Corpus [7], but this data is non-conversational and not comparable to the English data that has been used.

with others in the general Mixer population, but these subjects would speak in a room with a custom designed recording system that would simultaneously record their voices on eight channels including two cell phone headsets, a dictaphone, and five different microphones types resembling ones often found in courtrooms or interview rooms. These 105 participants could also make further telephone-only calls as part of the general Mixer collection.

4 2004 NIST Evaluation

The 2004 NIST Speaker Recognition Evaluation [14], taking place in March and April, will use some of the new Mixer data for its evaluation data set. It will therefore allow investigation of the effects of language, transmission type, and handset type on recognition performance. The multi-channel data to support investigation of the effect of the use of different microphone types on performance will not be available in time for this evaluation. The 2005 evaluation should include this data.

The evaluation is being designed for all trials to involve the use of different handsets (as indicated by the recorded phone numbers using caller identification) in the training and test segment data. Like the last several NIST evaluations, this one will include testing conditions with "extended" amounts of training data available for each target speaker, up to 16 entire conversation sides. (The core testing condition, required of all participants, will involve single conversation sides for both train and test data.) Therefore the frequently used handsets on which speakers receive calls will generally be used for training, while the unique handsets on which they initiate some calls will often be used for test. To the extent possible the multi-conversation side training data for a speaker will be drawn from a single handset and from conversations in a single language, but this will not always be possible when training consists of 8 or 16 conversation sides.

Tables 2, 3, and 4 provide some statistics (which are to be regarded as provisional) on the numbers of speakers and conversation sides from the Mixer collection used in this evaluation. Note that these figures are for conversation sides, while those in Table 1 are for whole conversations. Table 2 shows that sizable numbers of speakers and conversations will be included for each language, with a total of 304 different speakers being used. For over a hundred of these speakers, training with 16 conversation sides will be an option. For many of these it will also be possible to train multiple models using 8 (or fewer) conversation sides involving different handsets or different languages. The collection design results in larger numbers of other language conversation sides for training than for test but, as indicated in table 2, significant numbers of test sides in the other languages will be included.

Recent evaluations have shown the benefits for performance that may result from using word transcriptions provided by automatic speech recognition (ASR) systems when large amounts of training and test data are provided. This has been so even with ASR error rates as high as 50 percent. This year BBN has agreed to provide to all evaluation participants transcriptions generated by a relatively fast state-of-the-art system (similar to that described in [15]). This English recognizer will process all of the training and test data (including that in other languages). It will be of interest to see how much advantage this higher quality ASR system provides compared to the systems used in previous evaluations, and whether its "English" transcripts of speech in other languages proves to be of some use for speaker recognition.

Other Language Spoken	Speakers	Other Language Sides		English Sides	
		Train	Test	Train	Test
Arabic	51	294	98	370	138
Mandarin	46	241	62	280	154
Russian	48	275	65	331	147
Spanish	79	107	47	706	195
English only	84			895	285
Total	308	917	272	2582	919

Table 2: *Speakers included in the 2004 NIST evaluation by other language spoken and their numbers of training and test conversation sides in each language*

Type of Phone	Training Sides	Test Sides
Landline	1467	595
Cellular	849	366
Cordless	1164	222
Other/unknown	35	16

Table 3: *Phone transmission types of training and test conversation sides to be included in the NIST 2004 evaluation*

How Phone Used	Training Sides	Test Sides
Speakerphone	158	68
Headset	518	117
Ear-bud	184	64
Regular (hand-held)	2626	934
Other/unknown	29	16

Table 4: *Phone handset types of training and test conversation sides to be included in the NIST 2004 evaluation*

Table 3 shows that large numbers of landline, cellular and cordless conversation sides will be available in both the training and test data. And table 4 shows that large numbers of headphone and handheld sets will be included, with lesser numbers of speaker phone and ear-bud sets. It should be possible to obtain meaningful

results on how these factors, in either the training or test speech, or the match or mismatch between the two, affect recognition outcomes on a common set of speakers.

The 2004 evaluation will also offer for the first time an unsupervised adaptation option. The test segments to be run against each target speaker model will be ordered chronologically, and systems will have the option to use test segment data to update the model for the processing of subsequent segments against the model, without knowing whether or not the test segment contained the true target speaker. (The overall average ratio of target to non-target trials will be about one to ten.) Whether or not such adaptation is used by a system, it will also be possible to investigate how time differences in the collection of training and test data affect performance.

5 Future Plans

With the Mixer data collection by the LDC presently continuing, and only a minority of the speakers collected thus far included in the 2004 evaluation set, it is expected that this collection will be a rich resource for the evaluation in 2005 and perhaps beyond. The multi-channel collections should be included in these future evaluations.

The likelihood of securing data for two or more successive evaluations from a fixed data collection protocol will enhance the comparability of performance results across evaluations. While there has been clear progress over the course of the NIST evaluations over the past eight years, measuring this progress with significant precision is difficult because of changes in evaluation procedures, and most notably because of differences in the types of data that have been collected and used. As has been suggested, speaker recognition is exquisitely sensitive to differences in methods of speaker recruitment and telephone collection, so maintaining a fixed data collection procedure long enough to produce data for multiple evaluations is a valuable community service.

It should be noted that the NIST Speaker Recognition Evaluations are open to all research sites interested in this field and willing to participate and to report on their systems at the evaluation workshops.

6 Acknowledgement

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.

7 References

[1] J. P. Campbell and D. A. Reynolds, "Corpora for the Evaluation of Speaker Recognition Systems", *Proc. ICASSP '99*, Phoenix, Arizona, pp. 2247-2250

[2] Przybocki, M. and Martin, A., "NIST Speaker Recognition Evaluation – 1997", *Proc. RLA2C*, Avignon, France, April 1998, pp. 120-123

[3] Doddington, G., et al., "The NIST Speaker Recognition Evaluation – Overview, Methodology, Systems, Results, Perspective", *Speech Communication* 31 (2000), pp. 225-254

[4] Martin, A. and Przybocki, M., "The NIST 1999 Speaker Recognition Evaluation – An Overview", *Digital Signal Processing* 10, Num. 1-3, January/April/July 2000, pp. 1-18

[5] Martin, A. and Przybocki, M., "The NIST Speaker Recognition Evaluations: 1996-2001", *Proc. 2001: A Speaker Odyssey*, The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001, pp. 39-43

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

[7] J. Ortega-Garcia et al., "AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification", Proc. *ICASSP '98*, Vol. II, pp. 773-776

[8] Doddington, G., "Speaker Recognition based on Idiolectal Differences between Speakers", *Proc. Eurospeech '01*

[9] Nakasone, H. and Beck, S., "Forensic Automatic Speaker Recognition", *Proc. 2001: A Speaker Odyssey*, The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001, pp. 139-144

[10] "Effective, Affordable, Reusable Speech-to-Text (EARS)", *DARPA Information Processing Technology Office*, http://www.darpa.mil/ipto/programs/ears/

[11] Pallett, D., "A Look at NIST's Benchmark ASR Tests: Past, Present, and Future", *Proc 2003 IEEE Workshop on Automatic Speech Recognition and Understanding*

[12] Cieri, C., et al., "The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data", *Proc. 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, May 26-28, 2004

[13] Campbell, J., 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

[14] Martin, A. and Przybocki, M., "NIST Speaker Recognition Evaluation Chronicles", *Proc. Odyssey 2004*, The Speaker and Language Recognition Workshop, Toledo, Spain, May 31-June 3, 2004

[15] Schwartz, R., et al., "Speech Recognition in Multiple Languages and Domains: The 2003 BBN/LIMSI EARS System", *Proc. ICASSP 2004*