Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Effects of the New Testing Paradigm of the 2012 NIST Speaker Recognition Evaluation

Alvin F. Martin[1], Craig S. Greenberg[1], Vincent M. Stanford[1], John M. Howard[1], George R. Doddington[1], John J. Godfrey[2]

[1]National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
{alvin.martin|craig.greenberg|vincent.stanford|john.howard|george.doddington}@nist.gov,
godfrey.jack@gmail.com

Abstract

The 2012 NIST Speaker Recognition Evaluation was substantially different from the prior NIST speaker evaluations in its basic paradigm regarding system knowledge of most of the target speakers. This involved both a substantial increase in the amount of training data for most targets, and the provision of this data in advance of the evaluation with knowledge of these specific targets available to the system for all evaluation trials. We examine the performance effects of these changes, with contrasts provided by a limited number of targets with limited training not made known in advance and by one participant's system designed not to take advantage of the prior knowledge of multiple targets.

1. Introduction

The 2012 NIST Speaker Recognition Evaluation (SRE12) [1] featured a major change in the testing paradigm from that used in all of the prior NIST SRE's from 1996 to 2010. In the earlier evaluations the target speakers were unknown to the systems until the evaluation data was distributed. In most of these, the core test provided only a single session of training data for each target, typically a five to ten minute telephone conversation side. The evaluation rules carefully specified that the training data provided for each target could only be utilized in trials for which that target was the model speaker.

In SRE12, by contrast, most of the target speakers were specified months in advance of the evaluation period, and knowledge of these targets and their training data was allowed for speaker modeling and test segment scoring. These targets were speakers who had been utilized in prior evaluations, and systems were permitted to utilize all of their speech data that had been included in the corpora involved in the previous evaluations. For most, this included many telephone conversations of five to ten minutes each, and for many it included multiple interview sessions as well. Thus the amount of training data for most of the target speakers was considerably greater than what was available in prior evaluations.

This changed paradigm for SRE12 was motivated in part by scientific curiosity given the availability of all the prior evaluation data, but also by the recognition that, for some applications, the scenario of having large quantities of prior training data for a known set of speakers of interest can be a realistic one deserving of further investigation. It was a relatively straightforward undertaking to collect the ReMix Corpus of new telephone conversational data from some of the prior speakers, as described in the next section.

Section 2 provides some general information on SRE12 and the data it used. Section 3 examines primary system evaluation results with respect to the new paradigm. Section 4 looks at contrasting systems not taking advantage of the additional information offered by the new paradigm. Section 5 considers the performance effects of the amount of training data, or test data, provided for a speaker. Section 6 discusses the implications and possible plans for further evaluation.

2. SRE12 Overview

SRE12 was notably different from prior SRE's in several ways. The changes in how target speakers were defined and could be utilized in SRE12 are noted above. In addition, there was a change in the primary cost metric, test segments were included that contained environmental or (artificially) added noise, and systematic variation in test segment duration was explored. See [2] for further discussion of these latter changes.

There were approximately 1800 early release target speakers in SRE12, and around 70 released with the evaluation data. There was test segment data for only around 250 of the early release speakers; all the others were only used in non-target trials.

The core trial set of SRE12, required of all participants, consisted of about 1.8 million trials. Participants were also invited to submit results for an "extended" set of trials, involving the same speakers but using most of the possible non-target trial pairings, which included around 88.5 million trials.

Since most of the test speakers in the non-target trials were known in advance to the systems, but some were not, non-target trials could be of two types, involving either known or unknown speakers. The official scoring metric (cost function) in past NIST SRE's has been a linear combination of the miss rate (target trials) and the false alarm rate (non-target trials). For SRE12, it was appropriate to consider two false alarm rates, that for known non-target trials and that for unknown non-target trials. It was decided that the scoring metric should equally weight these two error rates in the core and extended tests, implying that for non-target trials the prior probability of a known speaker be taken as 0.5.

Participants were invited as well to submit results for contrasting systems, in addition to their primary systems, that always assumed that non-target speakers were unknown speakers (as in prior SRE's) or always assumed they were known speakers, in which case a single false alarm rate was utilized in the metric.

The previous evaluations have used a cost function weighting the false alarm rate at least 99 times that of the miss rate (more in SRE10), making the prior probability of a target trials 0.01 or less. For SRE12 it was decided to use weightings giving priors of 0.01 and of 0.001, and declare the official metric to be the mean of these two cost functions. (Using such an average made the calibration task with known and unknown non-targets more challenging.)

Five "common condition" subsets of the core trial set were selected for particular examination as described in [1,2]. These were trials involving multiple segment training and:

CC1: Interview speech in test without added noise
CC2: Telephone channel speech in test without added noise
CC3: Interview speech in test with added noise
CC4: Phone call speech in test with added noise
CC5: Phone call speech in test collected in a noisy environment

Table 1 indicates the numbers of core and extended trials of each type included for each of the common conditions.

Table 1: Numbers of core and extended trials for each common condition

Common Condition	Core (target / known non-target / unknown non-target)	Extended Trials (target / known non-target / unknown non-target)
1	2,897 / 46,601 / 61,871	3,860 / 10,985,377 / 11,349,426
2	7,354 / 445,041 / 105,196	7,354 / 10,312,118 / 2,088,834
3	3,851 / 49,032 / 20,048	5,127 / 12,444,672 / 4,804,500
4	7,176 / 411,843 / 4,872	7,176 / 9,471,219 / 124,830
5	3,883 / 209,532 / 2,406	3,883 / 5,119,130 / 77,745

Here we shall focus on first two common conditions, CC1 and CC2. An upcoming paper will address the effects of additive or environmental noise and of other performance factors as observed in SRE12.

The training data for the SRE12 previously known targets came from the several Mixer Corpora (Mixers 3, 4, 5, and 6) collected by the Linguistic Data Consortium and used in SRE08 and SRE10 (see [3,4,5]).

Two newer LDC corpora were the source of the SRE12 test segment data. One was Mixer 7, consisting of telephone-recorded phone calls, microphone-recorded phone calls, and microphone-recorded interviews collected for the IARPA Biometric Exploitation Science and Technology (BEST) Program [6,7]. The SRE06 target speakers not revealed in advance came from this corpus, with a single conversational telephone segment provided as training data.

The primary new corpus of test data for the targets revealed in advance was denoted ReMix. It consists (solely) of telephone-recorded phone calls of speakers included in one of the previous Mixer corpora. Each included speaker was encouraged to make twelve 10-minute new phone calls, with no two of them made on the same day. The target trials for the early release targets all came from this corpus.

3. Primary System Results

Performance results here are presented, as customary in NIST SRE's, as DET (Detection Error Tradeoff) curves [8] showing the system's possible operating points in terms of the resulting miss and false alarm rates. Note that for the core and extended test conditions there are two types of false alarm rates (known and unknown non-targets) involved, and thus the plotted false alarm rates in the DET curves are the means of the two. (It may be argued that, due to the calibration issues arising in this case, the use of DET is less meaningful than previously. See [9]. Results are shown for one leading system, but those for other leading systems are not dissimilar.

For one leading SRE12 primary system, Figure 1 (from [2]) shows performance on CC1 (Common Condition 1) when, for non-target trials, the test segment speaker is limited either to known or to unknown target speakers. Figure 2 (from [2]) provides a similar plot with respect to CC2 (Common Condition 2). The same full set of target trials is used in both curves of each plot.

It may be observed from these figures that for CC2, but not CC1, system performance was enhanced when non-target speakers were known. Known non-targets were more readily rejected than unknown.

The superior performance with respect to known non-targets for CC2 was perhaps to be expected, but the contrasting situation for CC1 may need some explanation.

The test segments used in SRE12 came from the ReMix and Mixer 7 Corpora as described in Section 2. ReMix was entirely telephone data, so the CC1 test segments all came from Mixer 7. As discussed above, the Mixer 7 known speakers were "known" via a single conversational phone segment. Thus the known non-target speakers for CC1 were known from only a single speech segment released at the time the evaluation data was released. Since CC1 involved microphone recorded interview test segments, these speakers were also "known" only under different channel conditions from that of the test segment. It is thus plausible that for CC1 prior knowledge of the non-target speaker proved to be of no value in terms of system performance.

In contrast, the known non-target speakers in the CC2 context were generally speakers known from prior SRE's with multiple

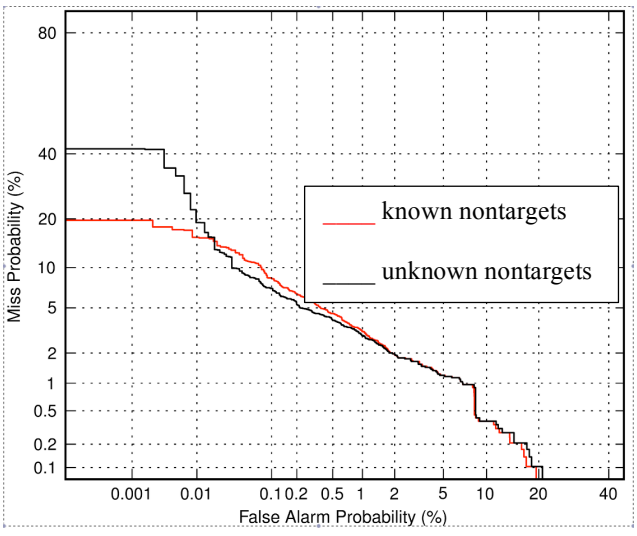

Figure 1: DET curves contrasting performance over known and unknown non-target speakers for one leading system for common condition 1.

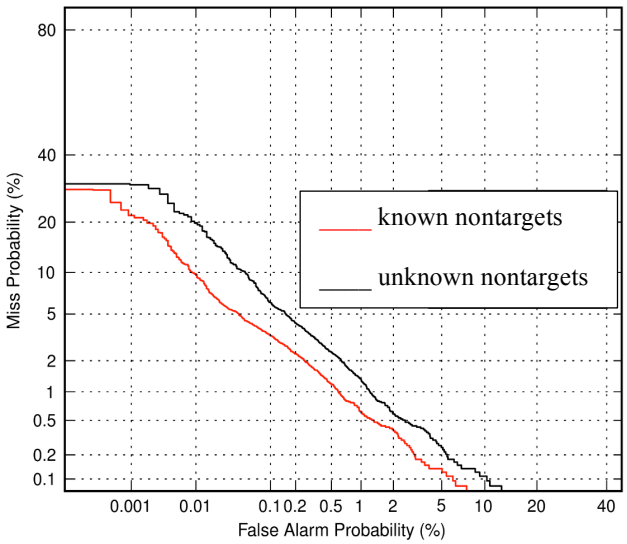

Figure 2: DET curves contrasting performance over known and unknown non-target speakers for CC2 for same system as in Figure 1.

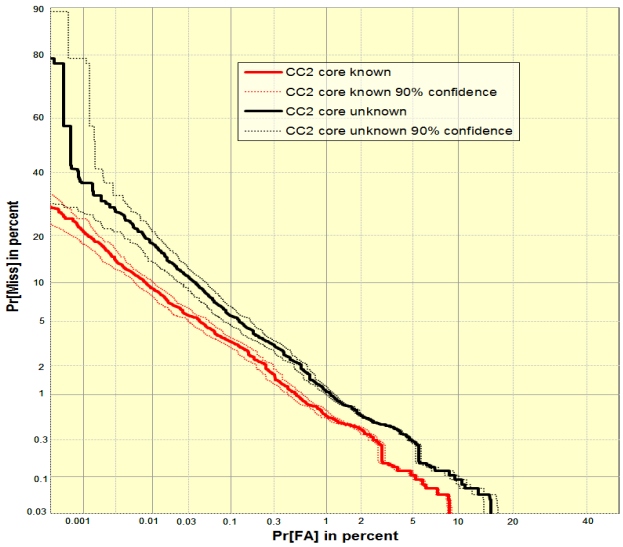

Figure 3: DET curves contrasting performance over known and unknown non-target speakers on extended trials satisfying CC2 for same participant as in Figure 1.

telephone conversations (and in some cases interviews) included in the training data provided in advance of the evaluation. The superior performance observed for these known speakers is an indication of the key effect of the changed evaluation paradigm.

The results presented are for one leading SRE12 system, but similar results have been observed for other leading systems.

How much confidence should we have in the significance of these results? The participant whose results are shown in Figures 1 and 2 also did the extended test. Figure 3, similar to Figure 2, shows its results on all extended trials satisfying Common Condition 2. In addition, it displays 90% confidence curves about the DET's. It thus supports the previous conclusions with reasonable confidence.

Note also in Figure 3 that the confidence bounds are much narrower at low miss rates and wider at high miss rates. This reflects the fact that the target trials are the same for both curves.

4. Contrasting Systems

The participating site involved in the preceding figures also submitted contrasting systems that assumed that all non-target speakers were known, and that assumed that all were unknown. Figure 4 shows similar plots as Figure 3 on trials satisfying CC2, for the contrasting system that assumed all non-targets trials involve unknown speakers. Here, as might be expected, there is little performance difference between known and unknown non-target speaker trials, with both curves similar to that for unknown speakers in Figure 3.

Figure 5 shows a similar plot of CC2 trials for the contrasting system where all non-target trials are assumed to involve known speakers. The overall contrast between performance on known and unknown speakers is similar to that in Figure 3, but the performance difference is a somewhat greater, as reasonably might be expected.

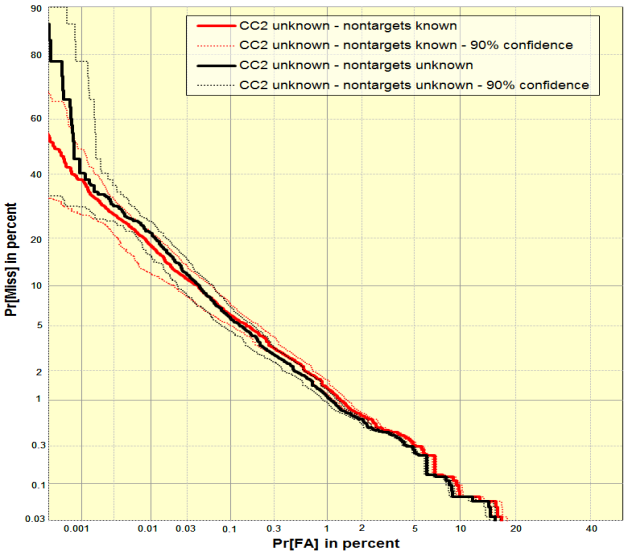

Figure 4: DET curves contrasting performance over known and unknown non-target speakers for CC2 extended trials for "unknown" system of the same evaluation participant as in Figure 1.

5. Amount of Training or Test Data

Figure 6 offers an indication of the impact on performance of the increased number of training sessions provided for most target speakers in SRE12. It compares performance, for one leading system (not the same as prior figures), on telephone channel trials when the number of training sessions was one versus that when the number of sessions was four or more.

SRE12 also examined in its core test the effect on performance of trial test segment duration, a factor not similarly included in the

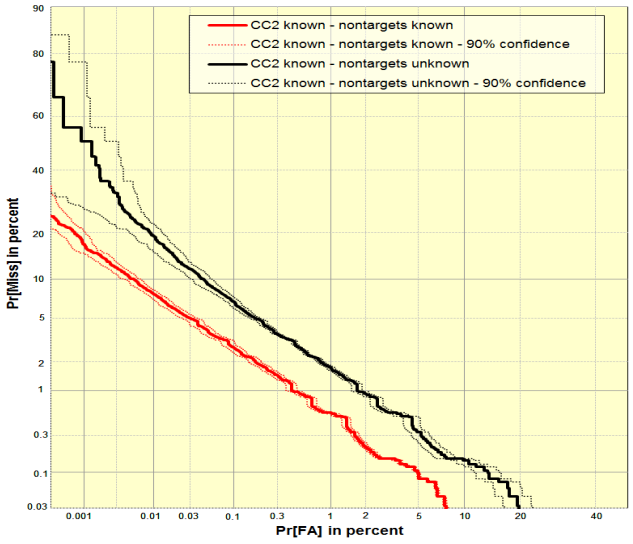

Figure 5: DET curves contrasting performance over known and unknown non-target speakers for CC2 extended trials for "known" system of the same evaluation participant as in Figure 1.

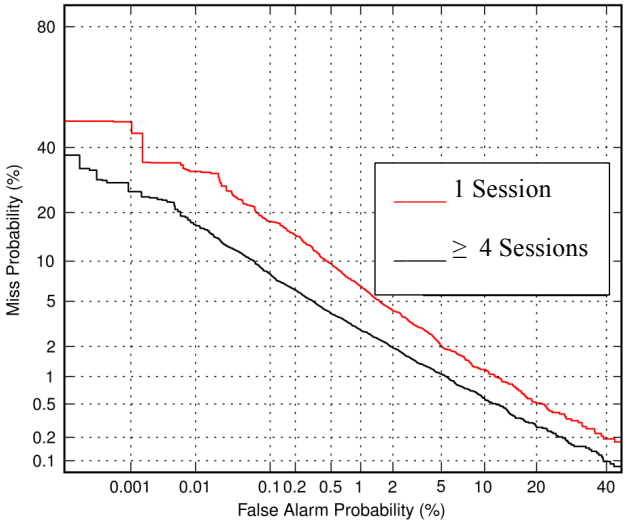

Figure 6: DET curves contrasting performance on telephone trials for one leading system when the target speaker training had only one session versus four or more sessions.

recent prior evaluations. Telephone test segments generally were approximately 30 or 100 or 300 seconds long. (On average, half of this contained speech from the target.) The previous CC2 plots included all durations. Figure 7 includes plots similar to those of Figure 3, involving the same system, but limited to 300 second segments. The dramatic performance improvement resulting from this may be observed. The effect of test segment duration on performance in SRE12 is further discussed in an upcoming paper.

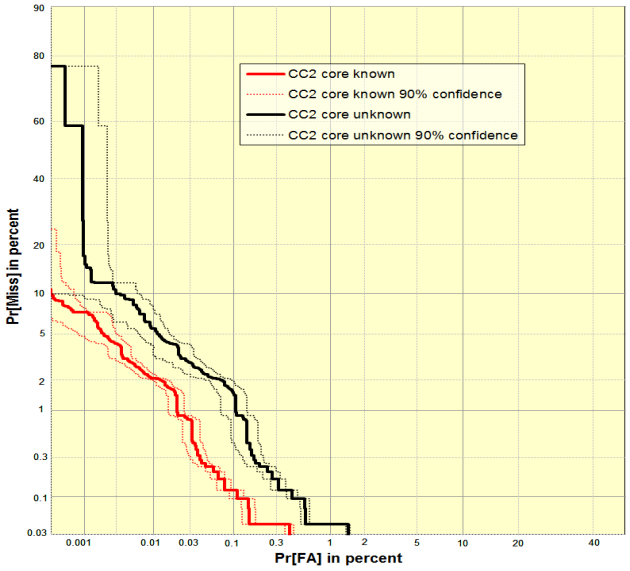

Figure 7: DET curves contrasting performance over known and unknown non-target speakers on extended trials satisfying CC2 but limited to the longer duration (300 s) test segments, for the same participant as in Figure 3.

6. Implications and Future Plans

The 2012 NIST Speaker Recognition Evaluation was an experiment with a different basic evaluation protocol in terms of how the target speakers were made known to the systems. With respect to conversational telephone speech, which has been the larger focus of the past NIST speaker evaluations, this change of protocol resulted in improved performance where known speakers were involved as anticipated. This was due both to the increased quantity of data available for the known speakers and to permitted knowledge of the non-targets for most trials.

The changed protocol was not implemented in a way that could be expected to show improved performance with respect to interview speech, and such improvement was indeed not observed.

This is perhaps a time for further discussion and consideration with regard to the protocols to follow in future evaluation. The issues involved are certainly very much application dependent. Many commercial applications can involve active and cooperative users who can only be expected to supply a limited amount of speech for enrollment. The NIST evaluations have focused more on having passive users for whom considerable speech, from one or many session, may be available for training. Here it is reasonable to expect targets to be known to the system, perhaps well in advance, via multiple training sessions, which could involve different types of speech.

The NIST evaluations have generally relied on other target speakers as the segment speakers in non-target trials. In actual applications, this is probably not a realistic situation. Thus if all or most targets are to be known to systems in advance, there is reason to want to have at least some of the non-target speakers not be among these.

The NIST speaker evaluations are expected to resume in the next couple of years. Effective evaluation depends upon the

collection of realistic and challenging speech data, and this is an expensive and time-consuming process. The next evaluation will be designed to take into account the lessons learned from the changed paradigm of SRE12.

7. Disclaimer

These results are not to be construed or represented as endorsements of any participant's system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials may be identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

8. References

[1] NIST, "The NIST Year 2012 Speaker Recogntition Evaluation Plan". Online: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

[2] C.S. Greenberg, et al., "The 2012 NIST Speaker Recognition Evaluation, *Proc. Interspeech 2013*, Lyon, France, Aug. 2013

[3] C. Cieri, et al., "Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4, and 5 Corpora", *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007

[4] L. Brandschain, C. Cieri, D. Graff, A. Neely and K. Walker, "Speaker Recognition: Building the Mixer 4 and 5 Corpora," in *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[5] L. Brandchain, et al., "The Mixer 6 Corpus: Resources for Cross-Channel and Text Independent Speaker Recognition", *Proc. LREC 2010*, Valletta, Malta, May 2010

[6] IARPA, "The Biometrics Exploitation Science & Technology Program". Online: http://www.iarpa.gov/Programs/sc/BEST/best.html

[7] C.S. Greenberg, A.F. Martin, and M.A. Przybocki, "The 2011 BEST Speaker Recognition Interim Assessment, Proc. Odyssey 2012, Singapore, Jun. 2012

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Eurospeech*, Rhodes, Greece, 1997, pp. 1899–1903.

[9] N. Brummer, "BOSARIS Toolkit - SRE'12 – How to deal with known non-targets". Online: https://sites.google.com/site/bosaristoolkit/sre12