The 2012 NIST Speaker Recognition Evaluation

Craig S. Greenberg[1], *Vincent M. Stanford*[1], *Alvin F. Martin*[1], *Meghana Yadagiri*[1],
George R. Doddington, John J. Godfrey[2], *Jaime Hernandez-Cordero*[2]

[1] National Institute of Standards and Technology
Gaithersburg, Maryland, USA
[2] US Department of Defense
Fort Meade, Maryland, USA

{craig.greenberg|vincent.stanford|alvin.martin|meghana.yadagiri}@nist.gov,
george.doddington@comcast.net, godfrey.jack@gmail.com, jherna2@tycho.ncsc.mil

Abstract

In 2012 NIST held the latest in an ongoing series of text-independent speaker recognition evaluations (SRE's). The 2012 NIST Speaker Recognition Evaluation (SRE12) was the largest and most complex SRE to date, including over 100 million trials. Several aspects of SRE12 were new; most significantly, NIST released in advance of the evaluation target speaker training data from six preexisting corpora, and systems were permitted to utilize joint information from target speakers for speaker modeling and test segment scoring. Results from the evaluation suggest that systems found it easier to reject non-target trials where the test speaker was among the target speakers.

Index Terms: speaker recognition, speaker detection, NIST SREs, standard reference data

1. Introduction

The 2012 NIST speaker recognition evaluation (SRE12) was the latest in an ongoing series that began in 1996. The task for this series has remained speaker detection: given training speech data from a target speaker, determine whether there is speech spoken by the target speaker in a test speech segment. The objective has been to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches[1]. This is accomplished by measuring system performance on common datasets using common metrics, enabling meaningful comparisons among approaches. The evaluations are intended to be of interest to all researchers working on the general problem of text independent speaker recognition; therefore they have been designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate[2].

There were several new aspects to SRE12, the most notable of which were that NIST released in advance of the evaluation target speaker training data from six preexisting corpora, and that systems were permitted to utilize joint information from target speakers for speaker modeling and test segment scoring. In addition, a large and variable amount of training data was available for each speaker, test segments with additive and environmental noise were included, system confidence scores were required to be expressed as log-likelihood ratios, and a new version of the cost metric was used. Unlike in prior SREs, automatic speech recognition transcripts for the evaluation test segments were not provided.

2. Data

2.1. Source Data

The target speakers for SRE12 consisted of the speakers used in SRE08 and SRE10 as well as some speakers not present in previous SRE's. The target speaker training data was drawn from six preexisting Mixer corpora collected by the Linguistic Data Consortium (LDC) in support of SRE's. These corpora consisted of telephone recorded phone calls, microphone recorded phone calls, and microphone recorded face-to-face interviews. More information on Mixer 1 and Mixer 2 can be found in [3], on Mixer 3, Mixer 4, and Mixer 5 in [4, 5], and on Mixer 6 in [6].

The evaluation test data was drawn from two corpora collected by the LDC that were unused in previous SRE's. The first, Mixer 7, was collected for the IARPA Biometric Exploitation Science and Technology (BEST) program [7], and consisted of telephone recorded phone calls, microphone recorded phone calls, and microphone recorded face-to-face interviews. The second, Remix, consisted solely of telephone recorded phone calls. The evaluation test segments were restricted to English language data. More information on Mixer 7 and the BEST speaker evaluation can be found in [8]. We briefly describe the Remix corpus below.

2.1.1. Remix

The Remix corpus consists of telephone recorded phone calls made by speakers who were included in at least one of the previous Mixer corpora.

Each speaker was encouraged to make twelve 10-minute phone calls, no two of which could be made on the same day. In addition, speakers were encouraged to make at least six of their phone calls in a noisy environment. Riding in a car or bus, having a television or music on in the background, or standing at a busy intersection were example noisy environments suggested to the Remix corpus speakers.

2.2. Evaluation Data Processing

The evaluation data was first multiplexed to be two-channel, with each channel consisting of one side of the conversation. Then segments of three different durations were selected from each two-channel recording. In addition, versions with added noise were also included for some of the segments.

Copyright © 2013 ISCA

2.2.1. Segment Selection

From each recording one segment of each of three durations was selected. These segments were nominally 30 seconds, 100 seconds, and 300 seconds in length. The first and last minute of each recording was removed prior to segment selection.

In order to select an n-second segment the recording was pre-emphasized and the root mean square (RMS) of each of the two channels was separately measured over an n-second window. The lesser RMS value of the two sides, what we'll call the channel min RMS, was associated with this window. This measurement was taken repeatedly as the window was moved across the pre-emphasized recording. The window with the maximum channel min RMS was selected in order to seek balance in the speech of the two channels.

2.2.2. Adding Noise

Noise was added only to the 300-second segments. The three noise types used were speech spectrum, HVAC (heating, ventilation, and air conditioning), and single speaker, with levels of 6 dB or 15 dB signal-to-noise ratio, using the ITU O.41 standard psophometric weighting (see [9] for details). Only a subset of the noise type and level combinations were applied to the data. Which noise and level combinations were applied was a function of whether the segment was an interview or telephone recorded phone call (no noise was added to microphone recorded phone calls). The noise type and level combinations used in SRE12 can be seen in Table 1.

Table 1: Noise types and levels used in SRE12

	6 dB	15 dB
HVAC	Interview	Interview
Crowd	Interview Telephone	Interview Telephone
Single Speaker	Telephone	

2.2.3. Data Encoding and Format

As in past evaluations, all test data were in SPHERE format and the telephone recordings were encoded as 8-bit μ-law sampled at 8 kHz. New to SRE12, all microphone channel data were encoded as 16-bit linear-pcm sampled at 16 kHz (rather than the prior 8-bit μ-law sampled at 8 kHz).

Beyond the information contained in a standard SPHERE header, the evaluation test segment headers included channel type (mic or tel) and speaking style (interview or phonecall).

3. Evaluation Conditions

SRE12 was divided into 10 distinct and separate tests. Each test involved one of three training conditions and one of six test conditions. One test was designated the core test, which all participants were required to complete. More information on the evaluation conditions may be found in [2].

3.1. Training Conditions

A training condition defines which data could be used to model target speakers. There were three training conditions in SRE12: *core, telephone,* and *microphone*. In the telephone (microphone) training condition, only data recorded over a telephone (microphone) channel could be used for creating target speaker models. There were no such training data restrictions in the core training condition.

3.2. Test Conditions

A test condition defines all conditions of an evaluation test outside of the training conditions. There were six test conditions in SRE12: *core, extended, supplemental, summed, known,* and *unknown*. These test conditions were characterized by three sub-conditions, namely test segment conditions, application parameter conditions, and trial selection conditions. These sub-conditions are described in the subsections below. Table 2 shows the sub-condition values for each of the test conditions.

Table 2: Sub-condition values for the six test conditions in SRE12

		Sub-Condition		
		Test Segment	**Application Parameter**	**Trial Selection**
Condition	**Core**	Two-channel	Mixed	Small
	Extended	Two-channel	Mixed	Large
	Summed	Summed Telephone	Mixed	Small
	Known	Two-channel	Known	Large
	Unknown	Two-channel	Unknown	Large
	Supplemental	Two-channel	Mixed	Large

3.2.1. Test Segment Conditions

A test segment condition describes the characteristics of the test data to be used in the trials. There were two test segment conditions in SRE12: *two-channel* and *summed telephone*. The two-channel test segment condition used two-channel test data recorded over microphone or telephone channels. The summed telephone test segment condition used single-channel test data consisting of sample-by-sample summing of the two channels.

3.2.2. Application Parameter Conditions

As mentioned above, for the first time in an SRE, systems were permitted to utilize joint information from target speakers for speaker modeling and test segment scoring. This created a meaningful distinction between those test speakers who were included among the target speakers ("known" speakers) and those who were not. The *a priori* probability that a test speaker is known was the application parameter. Its three values were *known, unknown, and mixed*, corresponding to respective *a priori* probabilities of 1.0, 0.0, and 0.5.

3.2.3. Trial Selection Conditions

A trial selection condition describes restrictions on the number the trials that may be included in the test. There were two trial selection conditions, *large*, and *small,* with limits of 100 million trials and 2 million trials respectively.

3.3. Training/Test Condition Combinations

The matrix of training and test condition combinations is shown in Table 3. Note that only 10 of the possible condition combinations were included in SRE12. The required (core) test, required of all participants, is highlighted.

3.4. Common Conditions

As in past SREs, subsets of trials with certain properties, announced in advance, were selected from the core test as common conditions of interest. Five common conditions were specified for SRE12. They consisted of all trials involving

Table 3: Matrix of training and test conditions. The shaded entry is the required core test

		Training Condition		
		Core	**Microphone**	**Telephone**
Test Condition	**Core**	required (core test)	optional	optional
	Extended	optional	optional	optional
	Summed	optional		
	Known	optional		
	Unknown	optional		
	Supplemental	optional		

multiple segment training and

1) interview speech in test without added noise
2) telephone channel speech in test without added noise
3) interview speech in test with added noise
4) telephone channel speech in test with added noise
5) telephone channel speech intentionally collected in a noisy environment in test without added noise

4. Performance Measures

The primary performance measure for SRE12 was a detection cost defined as a weighted sum of miss and false alarm error probabilities. Because the system confidence scores were required to be log likelihood ratios, the detection threshold was a known function of the cost parameters. Unlike in past SREs the primary cost measure was a combination of two costs, one using the cost parameters from SRE10 and one using a greater target prior. This change was suggested by members of the SRE community and was intended to add to the stability of the cost measure and to increase the importance of good score calibration over a wider range of log likelihoods.

In addition to the primary metric, C_{llr} scores were also computed. More information on C_{llr} can be found in [10].

4.1. Primary Metric

Equation (1) specifies the primary SRE12 cost function:

$$C_{Det} = C_{Miss} \times P_{Target} \times P_{Miss|Target} + C_{FalseAlarm} \times (1 - P_{Target}) \times (P_{FalseAlarm|KnownNonTarget} \times P_{Known} + P_{FalseAlarm|UnknownNonTarget} \times (1 - P_{Known})) \quad (1)$$

where the parameter C_{Miss} is the cost of a miss, $C_{FalseAlarm}$ is the cost of a false alarm, P_{Target} is the *a priori* probability that the segment speaker is the target speaker[1], and P_{Known} is the *a priori* probability that the non-target speaker is one of the evaluation target speakers[2]. The parameter values used in SRE12 are specified in Table 4.

To improve its intuitive meaning, C_{Det} was normalized by dividing it by the best cost that could be obtained without knowledge of the input data:

$$C_{Norm} = \frac{C_{Det}}{(C_{Miss} \times P_{Target})} \quad (2)$$

[1] Note that P_{Target} was not the same as the percentage of target trials among all trials.

[2] Note that P_{Known} was not the same as the percentage of known non-target trials among all non-target trials.

Table 4: Speaker Detection Cost Model Parameters

		C_{Miss}	$C_{FalseAlarm}$	$P_{Target-A1}$	$P_{Target-A2}$	P_{Known}
Test Condition	**Core**	1	1	0.01	0.001	0.5
	Extended					
	Summed					
	Known					1
	Unknown					0

The primary metric was calculated by computing C_{Norm} twice, once with P_{Target} equal to $P_{Target-A1}$ and once with it equal to $P_{Target-A2}$, and averaging the two values:

$$C_{Primary} = \left(C_{Norm}(P_{Target} = P_{Target-A1}) + C_{Norm}(P_{Target} = P_{Target-A2})\right)/2 \quad (3)$$

5. Results

Below we show results for primary systems on two of the common conditions describe in section 3.4. We also look at results for a leading system, comparing the performance for known and unknown non-target speakers. The results are displayed using DET curves. For more information on DET curves, see [11]. Note that here the false alarm rate plotted for each DET point represents the parenthesized expression in (1), and thus for the first three sub-conditions of Table 2 is the mean of the actual false alarm rates for known and for unknown non-target trials.

5.1. Common Conditions

Here we show the results for common condition 1 (multiple segment training and interview speech in test without added noise in test) and common condition 2 (multiple segment training and telephone channel speech in test without added noise in test). Performance curves for ten top systems are highlighted; others are in gray. It is important to note that the data making up the common conditions differ across the common conditions in ways unrelated to the common condition definition, e.g., the presence of environmental noise and relative balance of duration. Therefore performance differences observed across different common conditions cannot be used to infer that these differences are attributable to differences in the definition of the common conditions.

Figure 1 shows results for common condition 1. This condition involves interview segments, many of which feature

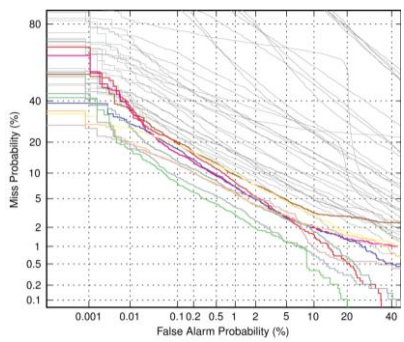

Figure 1: Primary System DET curves for Common Condition 1.

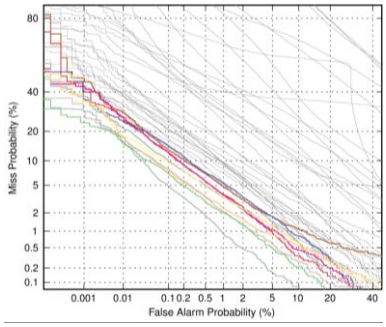

Figure 2: Primary System DET curves for Common Condition 2.

interviewer cross-talk on the subject channel, making the trials more challenging than they might otherwise be.

Figure 2 shows results for common condition 2. This condition involves telephone recordings, including those recorded in environmental noise.

5.2. Knowledge of Non-Target Speakers

One key area for SRE12 analysis is the comparison of known and unknown non-target speakers. This turns out to be affected by the channel and speech style of the test segments involved.

Figure 3 shows, for common conditions 1 and 2, plots contrasting DET curves of a leading system for known and unknown non-target speakers in the core test. (The target trials are fixed.) This figure suggests that for common condition 2, but not common condition 1, the system found it easier to reject known non-target speakers. This outcome held generally among leading systems.

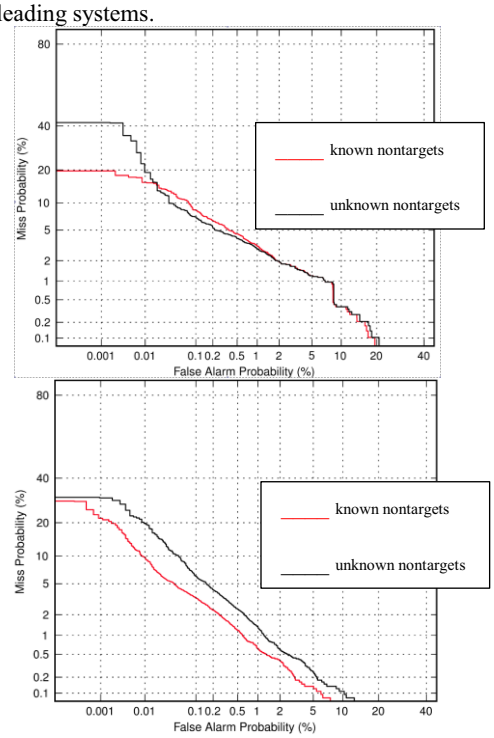

Figure 3: DET curves contrasting performance over known and unknown non-target speakers for one leading system for common condition 1 (above) and common condition 2 (below).

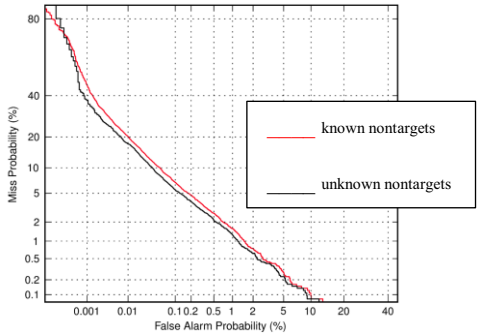

Figure 4: DET curves contrasting performance over known and unknown non-target speakers for common condition 2 for "unknown" system of the same evaluation participant as in Figure 3

This initially surprising result is explained by the fact that the interview test speakers are from the Mixer 7 Corpus, while most of the telephone test speakers are from Remix, and the training for the known Mixer 7 speakers consisted of only a single telephone test segment. Thus the common condition 1 known speakers were known only via one segment, and it was of a channel type and speaking style different from that used in test, thus a situation where misses were more likely.

Figure 4 provides a similar plot of common condition 2 trials for an “unknown” system from the same evaluation participant, i.e., a system treating all speakers other than the particular trial model speaker as unknown. This was the requirement for prior NIST evaluations, and the contrast for known speakers in the two common condition plots suggests the advantage conferred by the new SRE12 protocol.

6. Future Work

SRE12 was a large and complex evaluation with much opportunity for analysis. During the SRE12 evaluation analysis period prior to the workshop we looked at several factors affecting system performance, including noise, channel, and duration, and many others. We plan to publish details of these results in the near future. Even beyond this, however, much analysis remains to be done.

Plans for the next SRE are still in the early stages. We shall consider exploring further the sources of variability considered in SRE12 and additional sources of variability.

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 are 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, "NIST Speaker Recogntition Evaluations". Online: http://www.nist.gov/itl/iad/mig/sre.cfm

[2] NIST, "The NIST Year 2012 Speaker Recognition Evaluation Plan". Online: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf
[3] Cieri, C., Campbell, J. P., Nakasone, H., Miller, D. and Walker, K., "The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data," in *Fourth International Conference on Language Resources and Evaluation*, Lisbon, Spain, 2004.
[4] Cieri, C., Corson, L., Graff, D. and Walker, K. "Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora," in *Interspeech*, Antwerp, Belgium, 2007.
[5] Brandschain, L., Cieri, C., Graff, D., Neely, A. and Walker, K., "Speaker Recognition: Building the Mixer 4 and 5 Corpora," in *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.
[6] Brandschain, L., Graff, D., Cieri, C., Walker, K., Caruso, C., and Neely, A., "Mixer 6," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, 2010.
[7] IARPA, "The Biometrics Exploitation Science & Technology Program". Online: http://www.iarpa.gov/Programs/sc/BEST/best.html
[8] Greenberg, C.S., Matin, A.F. and Przybocki, M.A., "The 2011 BEST Speaker Recognition Interim Assessment," in *Odyssey Speaker and Language Recognition Workshop*, Singapore, 2012.
[9] ITU. (1995). *Psophometer for use on telephone-type circuits*. Online: http://www.itu.int/rec/T-REC-O.41-199410-I/en
[10] Brummer, N., and du Preez, J., "Application-independent Evaluation of Speaker Detection," in *Computer Speech and Language*, 2006, pp. 230-275.
[11] Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M., "The DET Curve in Assessment of Detection Task Performance," in *Eurospeech*, Rhodes, Greece, 1997, pp. 1899–1903.