Report on Performance Results in the NIST 2010 Speaker Recognition Evaluation

Craig S. Greenberg[1], *Alvin F. Martin*[1], *Bradford N. Barr*[1], *George R. Doddington*

[1] National Institute of Standards and Technology, Gaithersburg, Maryland, USA

craig.greenberg@nist.gov, alvin.martin@nist.gov, bradford.barr@nist.gov, george.doddington@comcast.net

Abstract

In the spring of 2010, the National Institute of Standards and Technology organized a Speaker Recognition Evaluation in which several factors believed to affect the performance of speaker recognition systems were explored. Among the factors considered in the evaluation were channel conditions, duration of training and test segments, number of training segments, and level of vocal effort. New cost function parameters emphasizing lower false alarm rates were used for two of the tests in the evaluation, and the reduction in false alarm rates exhibited by many of the systems suggests that the new measure may have helped to focus research on the low false alarm region of operation, which is important in many applications.

Index Terms: speaker recognition, speaker detection, NIST SREs

1. Introduction

The 2010 NIST Speaker Recognition Evaluation (SRE10) was the most recent in an ongoing series of speaker recognition evaluations conducted by NIST[1]. SRE10 used data from the Mixer-6 and Greybeard corpora collected by the Linguistic Data Consortium (LDC)[2][3]. New cost function parameters emphasizing lower false alarm rates were used for two of the tests in SRE10[4][5], including the core test of the evaluation, which all sites had to complete. Two extended tests were also conducted in which over six million trials were evaluated for each system brave enough to take on such a challenging task.

Section 2 briefly describes the Mixer 6 data utilized in SRE10. Section 3 shares evaluation performance results broken down by condition and compares SRE10 performance results with those of prior evaluations. System calibration is examined in Section 4. In Section 5 we discuss plans for future work.

2. Data

While SRE10 consisted of data from both the Mixer 6 and Greybeard corpora, the results presented in this paper are restricted to Mixer-6, which is now briefly described. For details, see [2].

2.1. Mixer-6

Mixer-6 collected speech over several microphones during interviews and phone calls conducted in two different recording rooms at the LDC. Telephone channel recordings were also collected for phone calls that took place outside the LDC. For each subject, two interviews and one phone call per day were collected over three non-consecutive days. Two of the phone calls that were recorded at the LDC were collected under conditions meant to induce either high vocal effort or low vocal effort from the subjects.

In order to induce high vocal effort, the subject wore aviator style headphones into which brown noise was mixed with the side tone and the interlocutor speech. The noise made it difficult for the subject to hear him- or her- self, inducing the Lombard effect [6].

Low vocal effort was induced by having the subject wear the same headphones, now not adding noise but raising the side tone level to make the subject hear him- or her- self very loudly.

3. Results

SRE10 utilized C_{Det}, described in section 4, as the primary evaluation metric. NIST also considers performance across a range of operating points and visualizes them using Detection Error Tradeoff (DET) curves, which are ROC curves with error rates on both axes plotted on a normal-deviate scale[7].

3.1. Vocal Effort

The Mixer 6 collection included audio recordings of low, normal, and high vocal effort phone calls. In order to test the effect of vocal effort on performance, trials involving training on normal vocal effort and testing on low, normal and high vocal effort were included in SRE10.

When training on normal vocal effort phone calls, testing on high vocal effort phone calls generally yielded worse results than testing on either normal vocal effort or low vocal effort calls. For nearly all systems, testing on low vocal effort phone calls yielded better performance than testing on normal vocal effort phone calls. This surprising result deserves further exploration.

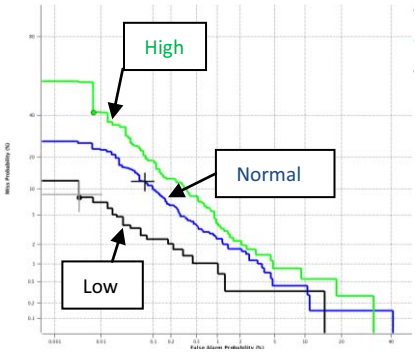

Figure 1: A leading system's DET plot, training on normal vocal effort phone calls and testing on low, normal, and high vocal effort phone calls. Note how low outperforms normal and high.

In order to compare the speech energy level for low, normal and high vocal effort, the average speech energy level for each session for each subject was computed. The distributions of this statistic for the three vocal effort conditions are shown in figure 2.

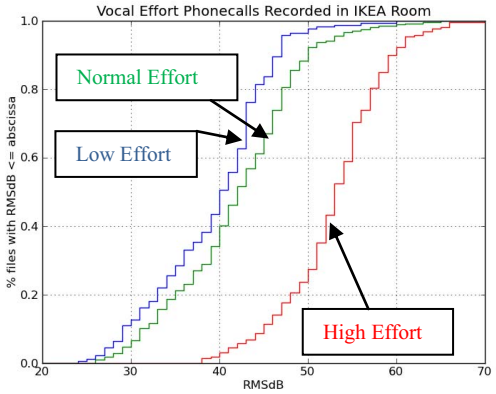

Figure 2: Cumulative histograms of RMS dB values for low, normal, and high vocal effort recordings in one of the two recording rooms.

These plots suggest that: the high vocal effort calls contained significantly more energy in the speech portions than did the normal or low vocal effort calls. Corresponding plots for the other recording room suggest the same conclusion and a comparison of the plots for the two rooms indicates that there was not a significant difference in the energy of the vocal effort recordings between rooms.

It is possible that natural variations in vocal effort between speakers affected these pooled histograms. So we conducted two paired tests to examine whether there was a significant difference in the energy of the speech when the speaker was fixed. The sign test found that the (positive) difference in energy between high and normal vocal effort was significant ($z = 17.104$), and that the low vocal effort did not differ significantly from normal vocal effort ($z = 0.1780$). The Wilcoxon signed-rank test gave the same results as the sign test: when comparing high and normal vocal effort, $w+ = 50{,}254 > 28{,}577$ (95% two-sided test) and when comparing normal and low vocal effort, $w+ = 20{,}709 < 22{,}950$ (95% two-sided test).

3.2. Channel Conditions

Fifteen microphone channels and one telephone channel were collected during Mixer 6 recording sessions. Eight of these channels, including telephone, were used in SRE10 (see Table 1). Only a subset of the possible channel combinations was tested (see Table 2).

Table 1: Channels from Mixer 6 used in SRE10

Ch. Num	Type	Dist. (cm) from speaker
2	lavaliere	20
4	podium	43
5	pzm	56
7	hanging	157
8	camcorder	71
12	shotgun	213
13	array	280

Table 2: The training and test channel combinations included in SRE10

Interview Train – Interview Test		Interview Train – Phone Call Test	
Train	Test	Train	Test
4	2		
4	4	4	4
4	5		
4	7		
4	8	4	8
4	12		
4	13		
		4	Telephone
8	7		
8	8	8	8
		8	Telephone

When training on interviews recorded over the podium microphone channel and testing on interviews, best performance was usually observed when the test segment was recorded over the lavaliere microphone channel or the podium microphone channel. The worst performance under these circumstances was usually when the test segments were recorded over the more distant hanging, shotgun, and array microphone channels. The performance for a leading system shown in Figure 3 illustrates this common trend.

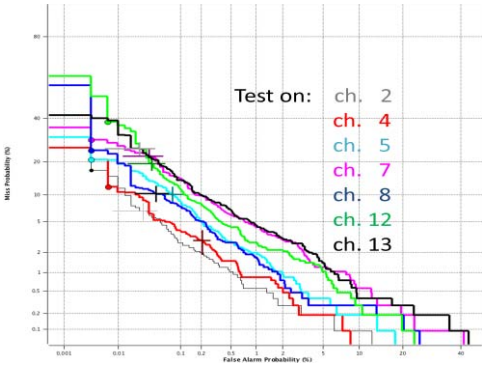

Figure 3: DET curves for a leading system when training on interview speech recorded over channel 4 and testing on interview speech

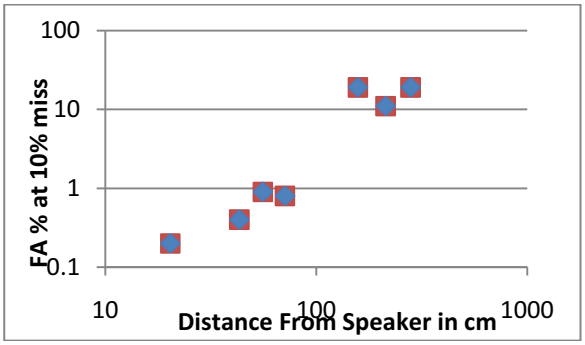

Figure 4: A scatter plot shows the change in false alarm rate at a 10% miss rate for a leading system as the distance between the speaker and microphone changes.

Figure 4 plots the false alarm percent at a miss rate of 10% for a leading system when training on interviews recorded over channel 4 and testing on interviews. This figure illustrates the major loss in performance when testing on far field microphones as opposed to near field microphones.

3.3. Duration of Training and Test Segments, and Number of Training Segments

The interview segments used in SRE10 were cut to either approximately eight minutes duration or approximately three minutes duration. Typically, training or testing using eight minute segments outperformed training and testing on three minute segments. See Figure 5.

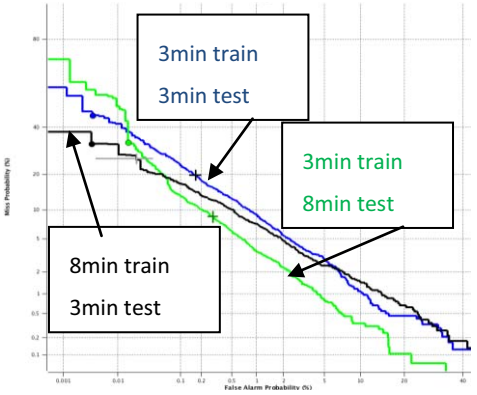

Figure 5: Performance of a leading system when training and testing on interview speech, varying the duration.

In addition to the core test, SRE10 included tests involving training on eight phone calls recorded over a telephone channel, each approximately five minutes in duration, as well as tests involving train and/or test segments with approximately 10-seconds of speech recorded over a telephone channel (see Figure 6). Similar to the varied interview durations in the core test, conditions with more speech exhibit better performance. It is interesting to note that systems performed better on the core test (train and test on one five-minute conversation) than when training on eight conversations and testing on ten seconds.

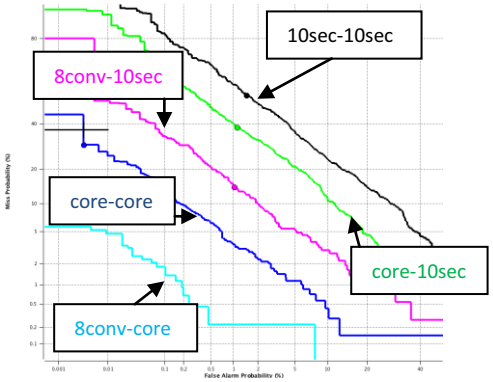

Figure 6: A leading system participating in several non-core tests in which multiple train segments were used and/or 10 second segments were used.

3.4. History

Figure 7 compares the performance of the top ystems in SRE06, SRE08, and SRE10 on trials involving English language telephone calls. Note the improvement in performance in SRE10 over the preceding two evaluations in the very low false alarm region.

Best system performance in SRE10 improved compared to SRE08 across the range of operating points when training and testing on interviews with mismatched microphones (see Figure 8).

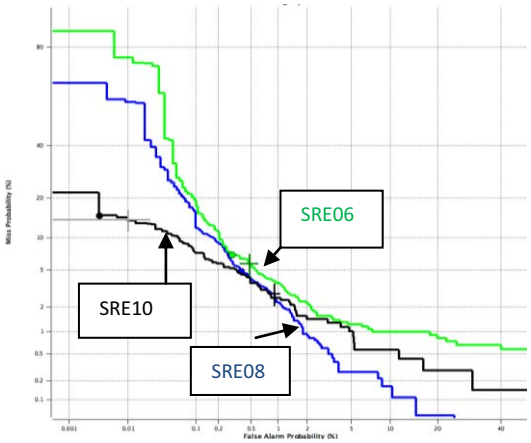

Figure 7: Best system performance on English telephone trials in SRE06, SRE08, and SRE10.

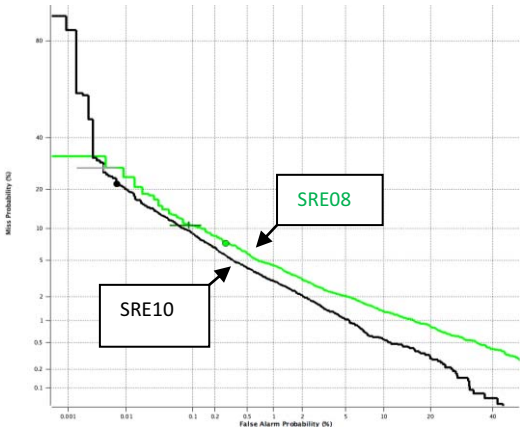

Figure 8: Best system performance when training and testing on interviews with mismatched microphones in SRE08, and SRE10

4. Calibration

NIST requires each system to output a "true" or "false" decision to the question, "is the speaker in the training segment(s) speaking in the test segment," for each trial and uses theses decisions to compute a detection cost metric. In addition, NIST requires a confidence score, where a higher score indicates a greater belief that "true" is the correct decision. Utilizing the system decisions, an actual cost is calculated. A minimum cost is computed by varying the system's decision threshold. The difference between the minimum and actual costs is a measure of system calibration. We here explore how well systems were calibrated in SRE10.

4.1. Evaluation Metric

The evaluation metric C_{Det} is defined as a weighted linear combination of miss and false alarm error probabilities:

$$C_{Det} = (C_{Miss} \times P_{Miss|Target} \times P_{Target}) + (C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1-P_{Target})),$$

where C_{Miss} is the cost of a miss, $C_{FalseAlarm}$ is the cost of a false alarm, P_{Target} is the chosen a-priori probability of a target trial, and $P_{Miss|Target}$ and $P_{FalseAlarm|NonTarget}$ are the observed miss and false alarm rates, respectively. SRE10 changed the parameters from preceding years for two tests, giving greater emphasis to performing at low false alarm rates, and the values for these parameters is given in Table 3.

Table 3: Original and new cost function parameters used in SRE10

	C_{Miss}	$C_{FalseAlarm}$	$P_{Miss|Target}$
Original	10	1	0.01
New	1	1	0.001

4.2. Calibration Performance

To measure a system's calibration, the ratio of actual cost to minimum cost averaged over the 9 common evaluation conditions is considered. (For details on the common conditions, see [4].). In Figure 9, these values are compared to the corresponding average minimum cost and average actual cost for each site. Note that several systems were well calibrated. Some systems with low minimum cost are off the chart in terms of actual cost, which is the evaluation's official metric. This is an indication that these systems, despite good recognition results, were poorly calibrated.

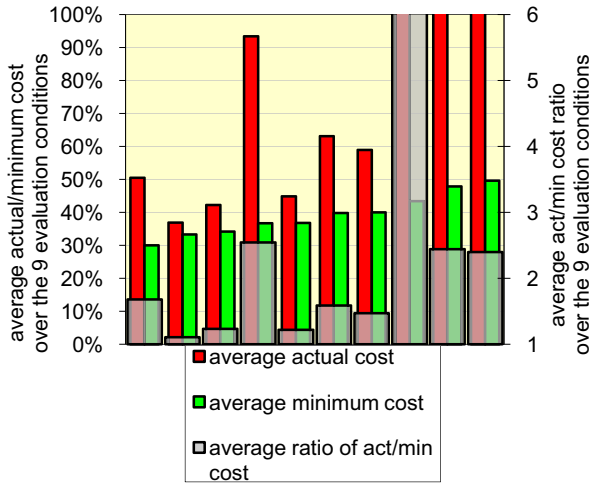

Figure 9: Comparison of average actual cost, average minimum cost and average actual/minimum cost ratio for systems with the best minimum cost.

5. Future Work

Despite a large amount of analysis, there remains much to explore in the SRE10 results. As mentioned in section 3.1, the analysis of the vocal effort results in particular deserves further exploration. Given the potential gains from further analyzing SRE10 results, NIST is planning to host an SRE11 analysis and review workshop in December of 2011, where further results on SRE10 data will be presented.

NIST plans on holding the next general Speaker Recognition Evaluation in the spring of 2012. As part of SRE10, NIST also ran a pilot evaluation of Human Assisted Speaker Recognition (HASR)[8] and is currently considering how to include HASR in SRE12.

6. 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.

7. References

[1] NIST, Information Technology Laboratory, "Speaker Recognition Evaluation", http://nist.gov/itl/iad/mig/sre.cfm.

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

[3] Brandschain, L., et al., "Greybeard – Voice and Aging", *Proc. LREC 2010*, Valletta, Malta, May 2010.

[4] A. F. Martin and C. S. Greenberg, "The NIST 2010 Speaker Recognition Evaluation", *Proc. Interspeech 2010*, Makuhari, Japan, September 2010.

[5] NIST, Information Technology Laboratory, "The NIST Year 2010 Speaker Recognition Evaluation Plan", http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf

[6] Lombard É, "Le signe de l'élévation de la voix", *Annales des Maladies de L'Oreille et du Larynx,* 37(2): 101–9, 1911.

[7] A. F. Martin, et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. Eurospeech '97,* Rhodes, Greece, September 1997, Vol. 4, pp. 1899-1903.

[8] C. S. Greenberg, et al., "Including Human Expertise in Speaker Recognition Systems: Report on a Pilot Evaluation", *Proc. ICASSP 2011*, Prague, Czech Republic, May 2011.