Human Assisted Speaker Recognition In NIST SRE10

Craig Greenberg, Alvin Martin*, Linda Brandschain[!], Joseph Campbell[#], Christopher Cieri[!], George Doddington, John Godfrey[^]*

*National Institute of Standards and Technology
Gaithersburg, Maryland, USA

[!]Linguistic Data Consortium
Philadelphia, Pennsylvania, USA

[#]MIT Lincoln Laboratory
Lexington, Massachusetts, USA

[^]US Department of Defense
Fort Meade, Maryland, USA

craig.greenberg@nist.gov, alvin.martin@nist.gov, brndschn@ldc.upenn.edu, jpc@ll.mit.edu, ccieri@ldc.upenn.edu, george.doddington@comcast.net, godfrey.jack@gmail.com

Abstract

The NIST series of Speaker Recognition Evaluations (SRE's) have, since 1996, evaluated automatic systems for speaker recognition. The 2010 evaluation (SRE10) also included a test of Human Assisted Speaker Recognition (HASR), in which systems based, in whole or in part, on human expertise were evaluated. Participants were invited to complete the trials in one of two small subsets of the full set of trials included in the core test of the main automatic system evaluation. The performance of these human dependent systems is currently being scored and analyzed. Their performance will be compared with the best automatic system results on the same trial subsets.

1. Introduction

NIST has coordinated evaluation of systems for automatic speaker recognition since 1996. See, for example, [1, 2, 3, 4, 5, 6, 7, 8].

A much discussed, but little tested, question about speaker recognition has been how much systems encompassing human expertise could add to speaker recognition performance. To address this matter, NIST offered in its 2010 Speaker Recognition Evaluation (SRE10) a test of Human Assisted Speaker Recognition (HASR). It consisted of two small sets of trials (denoted HASR1 and HASR2) that consisted of small subsets of the trials used in the core test of the primary evaluation of automatic systems. Participation in HASR was open to all interested sites utilizing systems involving, in whole or in part, human expertise and wishing to do either the HASR1 or HASR2 trials in accordance with the evaluation rules.

HASR is a new type of test for NIST evaluations, and accordingly is viewed as a pilot test. Depending on its outcome and the responses to it, it may be continued and refined in future evaluations.

Section 2 describes a preliminary experiment conducted prior to SRE10. The HASR rules and protocols are discussed in section 3, while the trial selection process is addressed in section 4. HASR participation is covered and some preliminary analysis of the HASR system results are presented in section 5. Section 6 concludes with a discussion of possible future plans for similar work.

2. Preliminary Experiment

To prepare for HASR trial selection, a preliminary experiment was carried out, involving the selection of a challenging set of speaker pairs for non-target trials and the presentation of these trials to a set of human trial evaluators.

The follow-up evaluation to the 2008 NIST Speaker Recognition Evaluation (SRE08) included a "full matrix" set of non-target trials involving all same-sex speaker pairs of the 150 speakers selected from the Mixer 5 Corpus [2]. The system submissions for these trials were examined to select 47 speaker pairs that resulted in errors by multiple different systems. All the interviews for each speaker in each pair were then listened to by humans and 10 apparently difficult to distinguish speaker pairs were selected.

A fifteen trial test set was then created that included 8 non-target trials involving 8 of the 10 selected pairs, along with 7 target trials in which the same speaker was used for both training and test. The non-target trials were selected from among the speaker pairs by listening in order to find utterances in which the speakers sounded most similar. The target trials were chosen from among the 52 unique speakers that were represented in the previously selected 47 speaker pairs by listening to pairs of segments from a given speaker and trying to find segments that sounded most different. In all cases, the recording channel was the subject's lavaliere microphone, generally the clearest channel available from the Mixer 5 Corpus.

A group of 14 other human evaluators was then asked to listen to these trials and decide whether each involved the same speaker or two different speakers. These participants were free to listen to the trials in whole or in part as many

times as they preferred, though some chose to limit the portions listened to or the number of repetitions.

The evaluators were generally volunteers with an interest in the project and some professional involvement in speaker recognition and its evaluation and applications. Yet this proved to be quite a challenging set of trials for these evaluators.

Figure 1 shows the total numbers of correct and incorrect decisions over all participants for the target and the non-target trials. The overall miss rate is over 18% and the overall false alarm rate of over 36%. The test did indeed prove to be a difficult one.

Figure 2 shows the numbers of errors, out of 14 decisions, for each of the 15 trials. Clearly some of the trials, both target and non-target were quite challenging, with three producing more errors than correct decisions, and three others error rates of a third or higher. Other trials were less difficult, with one target and one non-target trial having no errors.

Figure 3 indicates the numbers of errors on the 15 trials of each evaluator. Error counts are shown separately for target trials (out of 7) and non-target trials (out of 8). Notably, none of the evaluators had no errors, and half had total error rates of a third or more.

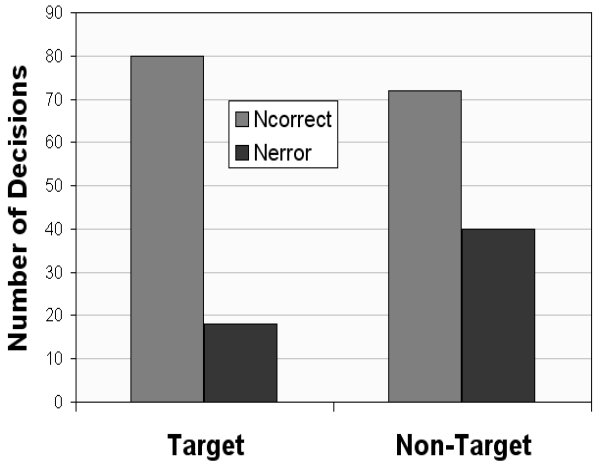

Figure 1: Numbers of errors and correct decisions for all submissions.

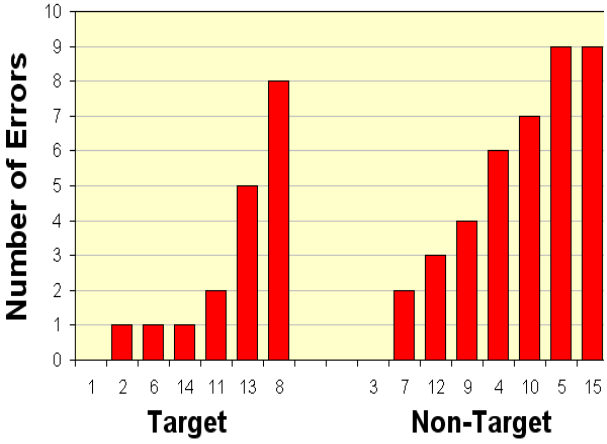

Figure 2: Numbers of errors (out of 14 decisions) for each of the 15 trials.

This experiment was viewed as supporting the belief that a meaningful HASR test set could be assembled even with as few as 15 trials, while recognizing that the statistical significance would be very limited. Many potential sites indicated reluctance to attempt more than this. It was decided however, to include in HASR both a 15 trial set (HASR1) and a superset of 150 trials (HASR2).

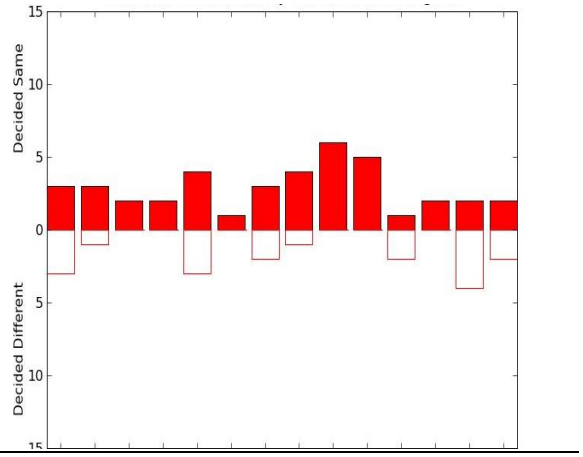

Figure 3: Numbers of errors (out of 15 trials) for each of the 14 evaluators (solid represents false alarms, outline represents misses).

3. HASR Rules and Protocols

The HASR rules and procedures were specified in the official 2010 evaluation plan [10].

The participating systems could be ones incorporating large amounts of automatic processing with human involvement in certain key aspects, or could be ones based solely on human listening, or something in between. The humans involved in a system's decision could be either a single person or a panel or team of people. These people might be professionals or experts in any type of speech or audio processing, or they might simply be "naïve" listeners. All participants were expected to provide system descriptions describing the people involved and both the automatic and human assisted processing algorithms utilized.

There could be interest in HASR from those involved with forensic applications of speaker recognition, and the participation of organizations in this area was welcome. The evaluation plan, however, notes the following:

> Forensic applications are among the applications that the HASR test serves to inform, but the HASR test should not be considered to be a true or representative "forensic" test. This is because many of the factors that influence speaker recognition performance and that are at play in forensic applications are controlled in the HASR test data, which are collected by the LDC following their collection protocols.

As in the main NIST evaluation, participants in HASR were expected to process each trial separately and independently of other trials. The specific rule was:

> Each decision is to be based only upon the specified test segment and target speaker model. Use of information about other test segments and/or other target speakers is not allowed.

This, however, represented a dilemma for human interactions, because humans inherently carry forward information from prior experience. To minimize this effect, a system was created to release the trials sequentially via an automatic email procedure. Each system was required to submit via email its result (decision and score) for each trial. Each trial submission was then automatically verified for correct format, following which the information needed to download the data for the subsequent trial was returned via automatic email.

The training and test speech data for each trial could be listened to by the human(s) involved in the processing as many times and in any order as might be desired. The procedures utilized and amount of human time devoted to this were expected to be documented in the system descriptions.

Participants were required to provide for each trial both a score and a decision, just as is required for the main evaluation of automatic systems. The decision was to be either "true" or "false", indicating whether or not the same speaker appeared in the training segment and the test segment. The score was to be a number, with higher scores indicating greater confidence that the two speakers were identical.

Because of the small numbers of trials, a simple approach to scoring was adopted. No cost function based on miss and false alarm rates, as used in the main SRE10 evaluation, was defined. Rather, for each system, the number of correct detections ($N_{correct}$ detections out of N_{target} trials) and the number of correct rejections ($N_{correct}$ out of $N_{non\text{-}target}$ trials) were to be reported.

While scores were required for each trial, it was recognized that where human judgments are involved, there may only a small discrete set of possible values. In the extreme, there might be only two scores; e.g., 1.0 corresponding to "true" decisions and -1.0 corresponding to false decisions. This was acceptable. For automatic systems in the main evaluation the scores are used to produce Decision Error Tradeoff (DET) Curves [11]; for HASR a discrete set of DET points will be generated.

4. Trial Selection

Past discussion with potential participants indicated disagreement about the number of trials that might reasonably be included. It was therefore decided to offer two HASR tests, with HASR1, consisting of just 15 trials and HASR2, consisting of 150 trials. The HASR1 trials were the first 15 of the 150 HASR2 trials. Participants could sign up for either HASR1 or HASR2.

Given the very limited number of trials included, particularly for HASR1, it was decided that special efforts should be made to include fairly difficult trials. This included both using an automatic system to identify a limited number of particularly challenging non-target trials, at least for such an automatic system, and listening to sub-select a smaller number of presumably more challenging trials for HASR1.

The HASR data was selected from the Mixer 6 corpus and, in all cases, training data were selected from interviews and test data were selected from phone calls. The trials, target and non-target, it should be noted, were selected to include varying recording channels, including ones involving training speech segments from interview sessions recorded over several different room microphone channels, and conversational telephone test segments recorded over a telephone channel, for some of which the subject wore a headset with voice feedback and noise levels selected to produce either high or low vocal effort.

Thus the HASR tests were expected to include at least some fairly hard trials. In hopes of finding challenging non-target trials for use in HASR1, potential non-target speaker pairs were selected by running an automatic speaker recognition system over a "full matrix" set of Mixer 6 interview segments. For each target speaker, we selected as a potential non-target speaker those who had at least six (of nine possible) false alarms in the top 1% highest scores of all trials run against the target speaker's models. This process yielded 37 speaker pairs. All interviews and phone calls in which a speaker present in the pairs participated were then listened to by humans. The segments that were perceived to be the most similar across speakers were chosen to be used in non-target trials.

This process was repeated to select the speaker pairs for HASR2, with the sole change being that only four false alarms were necessary for a non-target speaker to be considered instead of six. While this did yield some small number of cross sex potential speaker pairs, we chose to limit consideration to non-target speakers of the same sex as the target speakers. For HASR2, once the speaker pairs were chosen, the interview and phone call to be used for each pair were selected at random.

In order to find what were hoped to be challenging target trials for use in HASR1, the speakers represented in the 30 target trials with lowest scores in the automatic system output of the Mixer 6 interview full matrix were considered. All interviews and phone calls in which these speakers participated were listened to and what were perceived to be the least similar segments spoken by a single speaker were chosen as a target trial.

Target trials for HASR2 were selected by running an automatic speaker recognition system over a "full matrix" set of target trials, using models made from Mixer 6 interviews and Mixer 6 phone calls as test segments. The target trials with the lowest scores were selected for inclusion in HASR2.

Excerpts from each segment used in HASR2 trials were listened to in order to assure the presence of speech and that each trial would be relatively challenging. When, upon listening, a segment proved to be sufficiently anomalous in some way, e.g., lacking speech, the trial using the anomalous segment was removed and the next suitable trial, selected as described above, was used as a replacement.

5. Participants

Fourteen sites completed the 15 trials of HASR1. One of these sites ran three different systems on the trials, and another ran four different systems. Thus there were nineteen systems processing the HASR1 trials.

Five of the fourteen HASR1 sites completed the additional 135 trials included in HASR2, including the site with three different systems. Thus seven systems processed the HASR2 trials.

NIST has traditionally not publicly reported site names in association with their performance on speaker recognition evaluations. The possible sensitivity to public reporting by some sites could be of particular concern for a test such as

HASR. However, sites participating in HASR are welcome to report on their own performance in these tests.

6. Preliminary Analysis of Results

Presented here is some preliminary analysis of the HASR results. Further analysis and comparison with automatic system performance on the same trials remain to be carried out.

Figures 4, 5, and 6 provide information similar to that of Figures 1, 2, and 3 for the HASR1 trials. Figures 7 and 8 provide information similar to that of Figures 1 and 3 for the 135 HASR2 trials that are not included in HASR1.

Figure 4 shows the total numbers of correct and incorrect decisions across all systems on the HASR1 trials. The results correspond to a cumulative miss rate of almost 40%, and a cumulative false alarm rate in excess of 40%. These numbers are higher than for the human evaluators on the preliminary experiment, and seem quite surprising. The objective of choosing difficult trials appears to have been achieved.

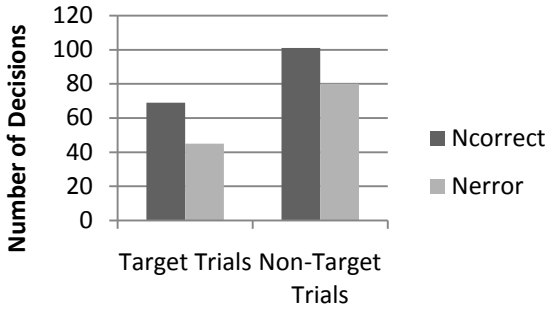

Figure 4: Total numbers of HASR1errors and correct decisions for target and non-target trials across all system submissions.

Figure 5 shows the numbers of the 19 systems in error on each of six target and nine non-target HASR1 trials. The results suggest that two of the target trials may have been less challenging than the others, while all of the non-target trials appear relatively challenging. It will be of considerable interest to examine, and listen to, the several trials for which more systems had an incorrect decision than a correct one

Figure 6 shows the numbers of misses (out of 6 target trials) and of false alarms (out of 9 non-target trials) on the HASR1 trials for each of the systems, plotted as error percentages on a DET plot. Scales for the actual numbers of misses and false alarms are added as well.

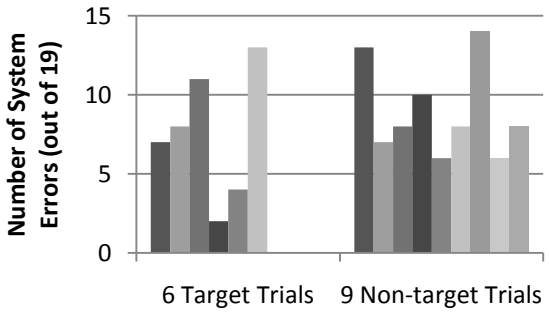

Figure 5: Numbers of errors (out of 19 systems) for each of the HASR1 trials

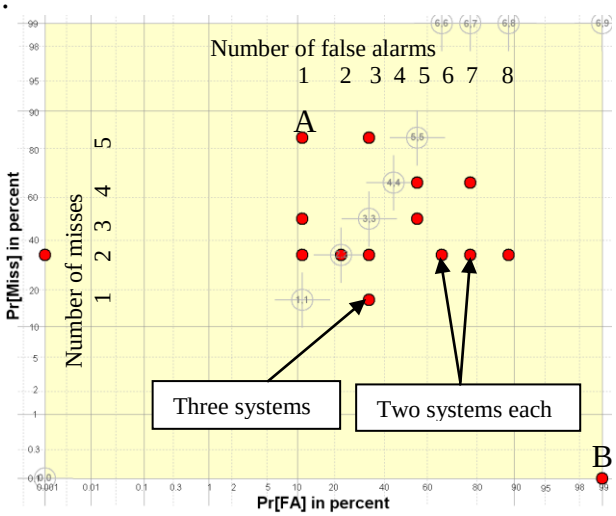

Figure 6: Numbers of misses and of false alarms of each of the 19 systems on HASR1 trials, plotted as DET points

Note that the multiple systems from two of the sites involve considerable similarities in their processing algorithms and the human expertise involved, and thus should not be viewed as independent samples of such systems. This should be considered with respect to all of these charts. That said, it may be noted that a majority of the systems (10 of 19) had more errors than correct decision on the 15 HASR1 trials. HASR1 was indeed a difficult test.

Figure 7 shows the total numbers of correct and incorrect decisions across all seven participating systems on the 135 HASR2 trials not included in HASR1. The results correspond to a cumulative miss rate of over 30%, and a cumulative false alarm rate in excess of 40%. These appear to be high overall error rates for trials expected to be somewhat less challenging than the HASR1 trials. These error rates are higher than those observed (with different humans involved) in the preliminary experiment.

Figure 8 shows the numbers of misses (out of 45 target trials) and of false alarms (out of 90 non-target trials) on these 135 trials for each of the seven HASR2 systems, plotted as error percentages on a DET plot. It may be observed in both Figure 8 and Figure 6 that different systems display operating points tuned to different tradeoffs between misses and false alarms. For example, the system points labeled A in figures 6 and 8 correspond to a common system with many misses but few false alarms, while the points labeled B in the figures correspond to a common system with few misses and many false alarms.

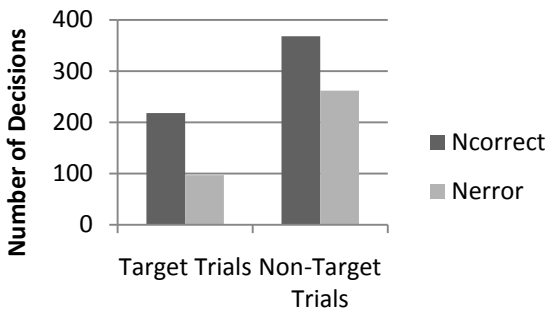

Figure 7: Total numbers of errors and correct decisions on the 135 HASR2 trials not included in HASR1 across all submitted systems

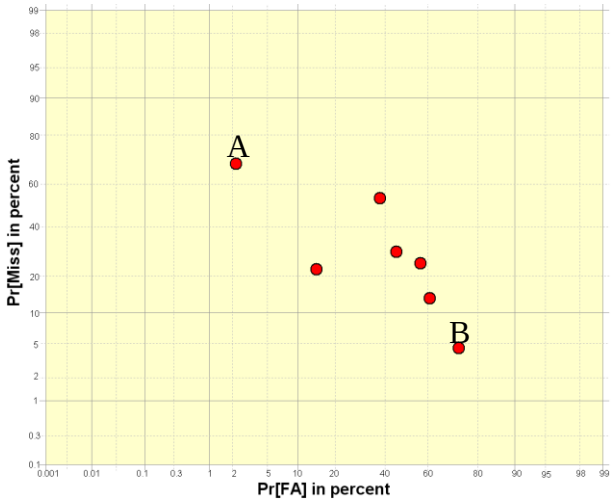

Figure 8: Numbers of misses and of false alarms of each of the 7 systems on HASR2 trials, plotted as DET points

The HASR submissions will be further analyzed, including examination of trial scores to determine their range of operating (DET) points and comparison with the best performing automatic systems of the main evaluation on the HASR trials. Limited statistical significance due to the small numbers of trials will very much constrain any inferences to be made, but this comparison could nonetheless prove very enlightening.

7. Future Plans

As noted, the HASR test included in SRE10 is considered a pilot. This pilot test and the results obtained will be discussed in detail at the SRE10 Workshop to be held in Brno, Czech Republic, June 24-25 of this year.

Further experiments on human speaker detection performance are being planned. A further test somewhat similar to the preliminary experiment described above is being organized with human volunteers. These trials will involve speech data from that used in SRE10 but with training and test segments of 30-seconds duration. Some trials will be similar to those used in HASR, while others will be chosen to be easier trials more typical of most SRE10 data. After another controlled test with known subjects, an experiment of offering such trials to anonymous, moderately paid subjects via the web service Mechanical Turk [12, 13] is also being planned. These experiments will also be discussed at the SRE10 Workshop.

Whether further formal HASR tests for automatic processing/human hybrid systems will be offered as part of future evaluations, or possibly as separate evaluations of their own, will depend on reviews of this pilot HASR test and related experiments.

8. 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.

9. References

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

[2] Reynolds, D. A., Keynote talk "Speaker and Language Recognition: A Guided Safari", *Proc. Odyssey 2008: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 2008

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

[4] Martin, A. F., Przybocki, M. A., and Campbell, J. P., "The NIST speaker recognition evaluation program", in Wayman, J. et al., editors, *Biometric Systems: Technology, Design and Performance Evaluation*, ch. 8, pp. 241-262, Springer, 2005

[5] Przybocki, M. A. and Martin, A. F., "NIST Speaker Recognition Evaluation Chronicles", *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop*, Toledo, Spain, June 2004

[6] Przybocki, , M. A., Martin, A. F., and Le, A. N., "NIST Speaker Recognition Evaluation Chronicles – Part 2", *Proc. Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, PR, June 2006

[7] Martin, A. F., "Evaluation of Automatic Speaker Classification Systems", in Muller, C, editor, *Speaker Classification I*, pp. 313-329

[8] Martin, A.F. and Greenberg, C.S, "NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels ", *Proc. Interspeech 2009*, Brighton, UK, September 2009

[9] Cieri, Christopher, Linda Corson, David Graff, Kevin Walker, "Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora", *Proc. Interspeech 2007*, Antwerp, August 2007

[10] NIST, Information , Technology Laboratory, Information Access Division, "2010 NIST Speaker Recognition Evaluation", *http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html*

[11] Martin, A. et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. EUROSPEECH-97*, Rhodos, Greece, pp. 1895-1898
[12] "Amazon Mechanical Turk", *http://aws.amazon.com /mturk/*
[13] Wikipedia, "Amazon Mechanical Turk", *http://en.wikipedia.org/wiki/Amazon_Mechanical_Turk*