Summary and Initial Results of the 2013-2014 Speaker Recognition i-vector Machine Learning Challenge

† *Désiré Bansé*[1*], *George R. Doddington*[1], *Daniel Garcia-Romero*[2], *John J. Godfrey*[2], *Craig S. Greenberg*[1], *Alvin F. Martin*[1], *Alan McCree*[2], *Mark Przybocki*[1], *Douglas A. Reynolds*[3]

[1] National Institute of Standards and Technology, Gaithersburg, MD
[2] Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD
[3] MIT Lincoln Laboratory, Lexington, MA
[*]Guest Researcher

desire.banse@nist.gov,george.doddington@comcast.net, dgromero@jhu.edu, godfrey.jack@gmail.com,craig.greenberg@nist.gov, alvin.martin@nist.gov, alan.mccree@jhu.edu, mark.przybocki@nist.gov, dar@ll.mit.edu

Abstract

During late-2013 through early-2014 NIST coordinated a special i-vector challenge based on data used in previous NIST Speaker Recognition Evaluations (SREs). Unlike evaluations in the SRE series, the i-vector challenge was run entirely online and used fixed-length feature vectors projected into a low-dimensional space (i-vectors) rather than audio recordings. These changes made the challenge more readily accessible, especially to participants from outside the audio processing field. Compared to the 2012 SRE, the i-vector challenge saw an increase in the number of participants by nearly a factor of two, and a two orders of magnitude increase in the number of systems submitted for evaluation. Initial results indicate the leading system achieved an approximate 37% improvement relative to the baseline system.

Index Terms: i-vector challenge, speaker recognition evaluation, SRE

10.21437/Interspeech.2014-86

1. Introduction

The National Institute of Standards and Technology (NIST) has been coordinating evaluations of speaker recognition technology since 1996 [1], the most recent of which occurred in late 2012 [2]. The task in the NIST Speaker Recognition Evaluations (SRE) has always been speaker detection, i.e., determine whether a specified speaker is speaking during a given segment of speech. The objective of this series is to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches.

Starting in December, 2013 and continuing through April, 2014, NIST coordinated a special i-vector challenge [3], based on the i-vector paradigm widely used by state-of-the-art speaker recognition systems [4]. Like the SRE series, the goal of the i-vector challenge was to foster research progress in order to improve the performance of speaker recognition technology. An additional goal of the challenge was to attract interest in the field from the machine learning community. Toward that end, the challenge was run entirely online and the data used were i-vectors rather than audio recordings. These changes made the challenge more readily accessible, especially to participants from outside the audio processing field. Using i-vectors in the challenge also enabled a comparison of system outputs using a consistent front-end and amount and type of training data.

† Authors appear in alphabetic order

In this paper we provide a description of the 2013 i-vector challenge task and an overview of the results obtained to date. In the next section we describe the task, data, experiment design, and performance metric which constitutes the i-vector challenge. In Section 3, we provide an overview of the participation rate and composition with results to date presented in Section 4. Finally, in Section 5 we discuss future directions for i-vector challenge series.

2. Evaluation Design

2.1. Task

The basic task evaluated in the i-vector challenge is speaker detection: determine if a particular person is speaking in an audio segment. In order to be evaluated, a system completes a set of trials, where a trial compares a target speaker model (defined by a set of training audio segments from a target speaker) to a test audio segment. A system must respond whether or not the speaker in the test segment is the target speaker by outputting a single (real) number, where a higher number indicates a greater degree of belief that the target speaker is the speaker in the test segment. The system's outputs are then compared with truth and a measure of performance for the system is computed.

To make this task more accessible to researchers outside the speech processing community, a single vector representation for each audio segment was supplied rather than providing audio segments.

2.2. I-vectors

While it is beyond the scope of this paper to describe in detail the i-vector approach (see [4] [5] for details) it is worth providing a high-level description. In the i-vector approach, an audio segment (e.g., one side of a telephone call) is first processed to find the locations of speech in the audio (speech activity detection) and to extract acoustic features that convey speaker information (typically mel-frequency cepstra and derivatives at 100 feature vectors/second). This sequence of feature vectors is then represented by their distribution relative to a Universal Background Model (UBM), which is a Gaussian mixture model (GMM) characterizing speaker-independent speech feature distributions. The parameters of this distribution are then transformed into a 600-dimensional vector using a total variability matrix T. This 600-dimensional vector is called an "i-vector". After normalizing the i-vectors,

a scoring function between a model and test i-vector is computed, the simplest version being the cosine distance between the average of i-vectors from the speaker's training segments and the i-vector representing the test segment [4]. The state of the art scoring function, called Probabilistic Linear Discriminant Analysis (PLDA) [5], requires a large collection (1000's of speakers each with 10's of i-vectors) of labeled training data.

2.3. Data

For the i-vector challenge, i-vectors from audio segments were supplied to focus research on ways to improve performance over the simple cosine distance baseline. These i-vectors were extracted using a speaker recognition system developed by the Johns Hopkins University Human Language Technology Center of Excellence in conjunction with MIT Lincoln Laboratory for the 2012 NIST Speaker Recognition Evaluation [6]. Standard MFCCs and deltas acoustic features and a GMM trained speech activity detector were used. The 2048 mixture UBM and T matrix used in i-vector extraction were trained using the development partition as described in [7]. Supplied with each i-vector was the duration of speech (in seconds) used to compute it. The segment durations were sampled using a log normal distribution with a mean of 39.58 seconds.

The data were organized into two datasets: a development dataset used for system creation, and a separate evaluation dataset for the challenge. The speakers involved in these two datasets were disjoint.

2.3.1. Development Data

The development data consisted of 36,572 i-vectors, provided for general system development purposes. These were from telephone call segments of unspecified speakers and could, for example, be used for unsupervised clustering in order to learn wanted and unwanted variability in the i-vector space.

2.3.2. Evaluation Data

The evaluation data consisted of sets of five i-vectors defining the target speaker models and of single i-vectors representing test segments. There were 1,306 target speaker models[1] (comprised of 6,530 i-vectors) and 9,634 test i-vectors.

The five i-vectors defining a given target speaker model were chosen from conversations utilizing a common telephone handset whenever possible. In some cases the handset used in a test segment matched that of a target model of the speaker.

2.3.3. Trials for Submission and Scoring

The trials of the challenge consisted of a mix of target and non-target trials. *Target trials* were those in which the target and test speaker are the same person. *Non-target trials* were those in which the target and test speaker were different persons. The full set of trials for the challenge consisted of all possible pairs involving a target speaker model and a single i-vector test segment. Thus the total number of trials was 12,582,004. It is worth noting that, unlike in the SREs, the challenge included cross-sex trials, which is a factor taken into account in our analysis of system performance.

[1] Note that there were instances where a single speaker had multiple target models.

The trials were divided into two subsets: a *progress subset*, and an *evaluation subset*. The progress subset comprised 40% of the trials and was used to monitor progress on a scoreboard viewable by all challenge participants. The remaining 60% of the trials formed the evaluation subset, which was used to generate the official, final scores at the end of the challenge. Which subset a trial belonged to was unknown to challenge participants, and each system submission had to contain outputs for all of trials.

2.4. Performance Measure

For each trial and each possible decision threshold value, an accept/reject decision was inferred according to whether the system's trial output score exceeded the threshold or not. When reject was decided for a target trial, this was a miss error; when accept was decided for a non-target trial, this was a false-alarm error. By using the sorted values of outputs from a system as thresholds, the system's miss and false-alarm rates were determined at all possible a-posteriori thresholds.

The overall performance measure was based on a *decision cost function* (DCF) representing a linear combination of the miss and false alarm error rates at a threshold:

$$\mathrm{DCF}(\mathit{thresh}{=}t) = (\#\,\mathrm{misses}(\mathit{thresh}{=}t) \,/\, \#\,\mathrm{target\ trials}) + (100 \times \#\,\mathrm{false\ alarms}(\mathit{thresh}{=}t) \,/\, \#\,\mathrm{non\text{-}target\ trials})$$

The minimum DCF obtained over all threshold values was the official system score recorded for a submission. This represents a change from the SRE series, where systems are required to choose the threshold used to calculate the official system score. Thus for each system submission, the performance score returned during the challenge was this minimum DCF over the set of trials used for progress scoring. At the conclusion of the challenge, the score for each participant's final submission was determined based on the evaluation subset of trials.

2.5. Baseline System

A baseline system was distributed to challenge participants in order to serve as an example of how to achieve a successful submission. The algorithm used in the baseline was a variant of cosine scoring. Unlike in the typical supervised setup for cosine scoring [4], WCCN and LDA were not used due to the lack of speaker labels in the development dataset; instead, the i-vectors were pre-processed using an unsupervised technique that performed centering and whitening based on the statistics of the development data, as follows:

1. Use the unlabeled development data to estimate a global mean and covariance.
2. Center and whiten the evaluation i-vectors based on the computed mean and variance.
3. Project all the i-vectors into the unit sphere.
4. For each model, average its five i-vectors and then project the resulting average-model i-vector into the unit sphere.
5. Compute the inner product between all the average-model i-vectors and test i-vectors.

2.6. System Requirements

Each uploaded system submission was required to contain outputs for all trials in order to be scored. The output produced

for each trial was required to be based solely on the training and test segment i-vectors provided for the trial (along with the development data). Use of any of the i-vectors provided for other trials was not permitted. For example, the following were not allowed:

- Normalization over multiple test segments
- Normalization over multiple target speakers
- Training system parameters using data not provided as part of the challenge
- Use of evaluation data for impostor modeling

3. Participants

As mentioned in section 1, the evaluation was run entirely online. The online platform was launched on December 04, 2013 and was taken offline on June 30th, 2014 (though the deadline to submit official results for the challenge was April 7th, 2014). A total of 259 participants registered from 47 different countries. The greatest number of participants were from the USA (60), China (28), and India (18). Figure 1 displays the number of participants from each country. It should be noted that all participant information, including country, was self-reported. Of the 259 participants, 109 participants, representing 100 unique sites, submitted at least one valid submission. This number exceeds the number of sites in the SRE12 by nearly a factor of 2.

Figure 2 shows the relative number of participants per affiliation type. Most participants came from academia, and many others came from industry or reported being self-affiliated.

In order to reduce the number of systems differing only with respect to small changes in hyper-parameters, the SRE series has generally limited the maximum number of submissions per site to 3 systems total. The i-vector challenge on the other hand set a daily limit of 10 submissions for each participant. This was done for logistical reasons, namely to limit the number of submissions NIST would need to score at any given time.

During the official scoring period challenge participants submitted in excess of 5900 submissions. Figure 3 shows the numbers of submissions participant made during the evaluation period. Participants made on average between 4 and 5 submissions (mean = 4.395, standard deviation = 3.122) each day. One participant submitted more than 500 system outputs during the evaluation period!

Figure 1: Bubble plot showing participants per country (larger circles indicate a greater number of participants).

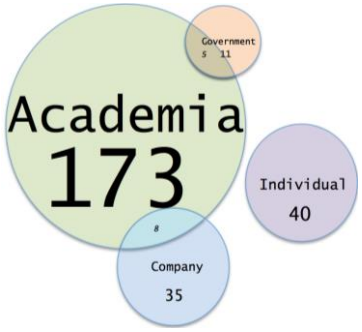

Figure 2: Relative numbers of participants for each affiliation type. Larger circles (and intersections) represent larger numbers of participants.

Table 1 compares the number of participants and system submissions between SRE12 and the i-vector challenge. Both the total number of participants and the number of new participants grew by an approximate factor of 2 in the i-vector challenge. The number of systems submitted for the i-vector challenge was two orders of magnitude greater than in SRE12, which increased the challenge of successfully scoring and analyzing all of the submitted systems. These increases in participation suggest that the i-vector challenge was successful in reducing the barrier of participation.

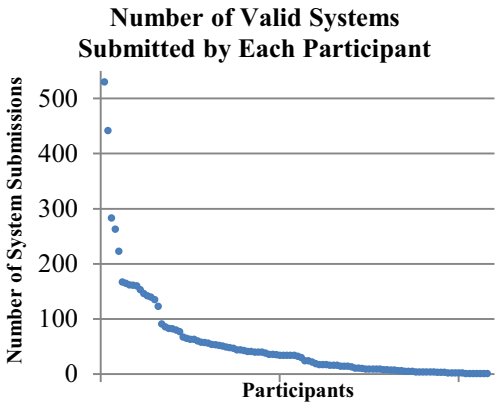

Figure 3: Number of systems submitted by individual participant.

Table 1: A comparison of participation between SRE12 and the i-vector challenge

	SRE12	i-vector 2014
# of Sites	58	100
# of New Sites	16	36
# of System Submissions	212	5900

4. Initial Results

It should be noted that the results reported below are as of the time of this paper's submission and are likely to change before the challenge closes. They are also limited to the progress set; initial results on the evaluation set can be found in [7].

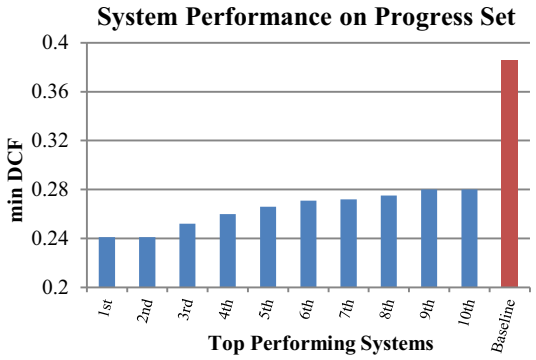

Figure 4: The min DCF for the ten top performing systems

In Figure 4 we see the min DCF values on the progress set for the baseline system as well as the best performing systems for the 10 leading participants. At the end of the official scoring period, the baseline system, with a min DCF value of 0.386, ranked 85th out of 110 (109 participants and 1 baseline system) on the progress set, meaning approximately 80% of challenge participants submitted a system that outperformed the baseline. The leading system at the end of the evaluation period had a min DCF value of .241, which represents an approximate 37% relative improvement over the baseline. The participant with the 10th lowest min DCF had a value of .280, an approximate 27% improvement over the baseline. Of the 5,900 systems submitted, only 36% performed better than the baseline system.

In Figure 5 we see the lowest min DCF value at any given time over all submitted systems, along with the min DCF value at that time for the participant submitting the system with the leading performance on the progress set at the end of the evaluation period. It is worth noting that the relative rate of performance improvement decreased rapidly, with little improvement observed after the evaluation had been running for 6 weeks.

Unlike in the SRE series, the i-vector challenge included trials where the target speaker was a different gender than the model speaker (i.e., cross-sex trials). Figure 6 shows the performance of the baseline system on male only, female only, same-sex only, and all trials. There is a relative performance improvement of approximately 13% on all trials (including cross-sex trials) relative to same-sex only trials.

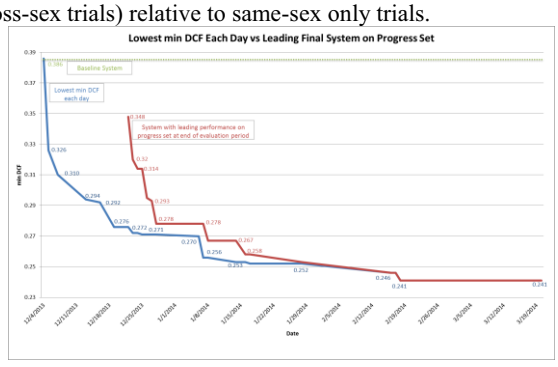

Figure 5: Min DCF on the progress set over time. The blue line shows the lowest min DCF on a given day. The red line shows the min DCF for the participant with the leading performance on the progress set at the end of the evaluation period.

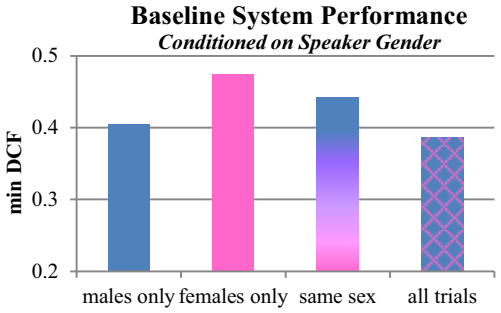

Figure 6: Baseline system performance on the progress set for male-only, female only, same-sex only, and all trials.

5. Discussion and Future Work

When comparing systems in the SRE series, it is often difficult to determine without further experiment whether performance differences are primarily due to changes in the data preparation or changes in the modeling approach. One of the advantages of having utilized i-vectors rather than audio as the input data for this evaluation was that data preparation was fixed for all participants. Thus observed performance differences could immediately be attributed to changes in modeling approach. Another advantage was that participants without a strong background in audio processing could participate in the challenge more easily. We were pleased with the increased participation over SRE12, particularly the inclusion of many new participants, though it was difficult to determine how much new participation was enabled specifically by the change from distributing audio to distributing i-vectors.

The i-vector challenge was the first evaluation of speaker recognition technology that NIST coordinated entirely online, thus a new platform was built to support it. Despite some initial technical challenges and administrative burdens to get the platform online, from our perspective the change was positive and, by and large, smooth. We made several platform enhancements during the evaluation period and are planning several more; for example we would like to make performance analysis tools available through the platforms. We anticipate running future challenges in a similar manner, and expect future SREs to incorporate elements of the i-vector platform.

6. Disclaimers

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.

Lincoln Laboratory's work was sponsored by the Department of Defense under Air Force contract FA8721-05-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] NIST, "NIST SRE Homepage," [Online]. Available: http://www.nist.gov/itl/iad/mig/sre.cfm.

[2] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey and J. Hernandez-Cordero, "The 2012 NIST Speaker Recognition Evaluation," in *Interspeech*, Lyon, France, 2013.

[3] NIST, "NIST i-vector Challenge Homepage," [Online]. Available: http://www.nist.gov/itl/iad/mig/ivec.cfm.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 19, no. 4, pp. 788-798, 2011.

[5] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, Florence, Italy, 2011.

[6] NIST, "NIST SRE12 Homepage," [Online]. Available: http://www.nist.gov/itl/iad/mig/sre12.cfm.

[7] D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, C. S. Greenberg, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki and D. A. Reynolds, "The 2013-2014 Speaker Recognition i-vector Machine Learning Challenge," in *Odyssey*, Joensuu, Finland, 2014.