Analysis of the Second Phase of the 2013–2014 i-Vector Machine Learning Challenge

Désiré Bansé1, George R. Doddington1, Daniel Garcia-Romero2, John J. Godfrey2, Craig S. Greenberg1, Jaime Hernández-Cordero4, John M. Howard1, Alvin F. Martin1, Lisa P. Mason4, Alan McCree2, Douglas A. Reynolds3

1 The National Institute of Standards and Technology, Gaithersburg, Maryland USA
2 Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, Maryland USA
3 MIT Lincoln Laboratory, Lexington, Massachusetts USA
4 U.S. Department of Defense, Fort Meade, Maryland USA

desire.banse@nist.gov, george.doddington@comcast.net, dgromero@jhu.edu, godfrey.jack@gmail.com, craig.greenberg@nist.gov, jherna2@tycho.ncsc.mil, john.howard@nist.gov, alvin.martin@nist.gov, lpmicke@tycho.ncsc.mil, alan.mccree@jhu.edu, dar@ll.mit.edu

Abstract

In late 2013 and 2014, the National Institute of Standards and Technology (NIST) coordinated an i-vector challenge utilizing data from past NIST Speaker Recognition Evaluations. Following the evaluation period, a second phase of the challenge was held, where speaker labels were made available for system development. The second phase included system submissions from 23 participants representing 13 different countries, of which 18 also participated in the first phase of the challenge. The top 10 systems participating in both of the challenge phases demonstrated an average relative improvement of approximately 26% between the first and second phases, which represents the value of having access to the speaker labels. The top five participants submitted a system that outperformed the oracle system from the first phase on the evaluation data.

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

1. Introduction

From December 2013 to April 2014, the National Institute of Standards and Technology (NIST) coordinated an i-vector challenge utilizing data from past NIST Speaker Recognition Evaluations [1]. This challenge was intended to foster interest in the speaker recognition field from the machine learning community and used low-dimensional feature representations (i-vectors) rather than audio in order to make the challenge accessible to participants without an audio processing background. For more information regarding i-vectors see [2].

The challenge was posed initially as an unsupervised learning task. In particular, all data available for system development was unlabeled, so systems did not know which development i-vectors came from a common speaker. (Development speakers were disjoint from those used in the test data.) At test time, the systems were given sets of i-vectors and were required to output a single (real) number, where higher numbers indicate greater degree of belief that the i-vectors were extracted from audio containing speech from a single speaker. For a general summary of the results, see [3].

The most common approaches taken by participants used clustering methods. See, for example, [4, 5, 6, and 7] for discussions of some of the various approaches that were undertaken. The leading system achieved a relative improvement of approximately 38% over the fairly simple baseline system that was provided. These results were also compared to an “oracle” system, one that had access to the development data labels, and was a sophisticated system near the state of the art. For more on related results, see [3].

The i-vector challenge evaluation period ended in April 2014, after which a decision was made to conduct a second phase of the challenge, where systems were given access to development speaker labels at train time, making it a supervised learning task. This paper reports on results of this second phase of the challenge, which ran from July 22nd, 2014 through September 15th, 2014.

2. Data

The second phase included speaker labels for the training data. Otherwise, the second phase data was identical to that of the first phase.

Briefly, the data provided for system development consisted of 36,572 i-vectors. The evaluation data included 6,560 i-vectors forming 1,306 sets of 5 i-vectors - defining the target speaker models - and 9,634 test i-vectors. All of the provided i-vectors were 600-dimensional. See [3] for more details.

3. Participation

Table 1 compares the participation in SRE12 with that of the first and second phases of the i-vector challenge. There were...
105 sites that participated in the first phase of the challenge, exceeding expectations. Participation remained high during the second phase, with 46 participants downloading the updated dataset, half of whom submitted at least one system. Out of these, 25 were new to the challenge. Among the participants submitting at least one system in the second phase, five were new and had not submitted in the first phase. A total of 1,431 system outputs were submitted during the second phase of the challenge and, counting some submissions between phases, the total outputs exceeded 10,000. Thus submissions for the i-Vector Challenge exceeded those for LRE12 by a factor of about 50.

### Table 1: A comparison of participation between SRE12, i-Vector Challenge Phase 1, and Phase 2.

<table>
<thead>
<tr>
<th>Metric</th>
<th>SRE12</th>
<th>Phase 1</th>
<th>Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># Sites</td>
<td>58</td>
<td>105</td>
<td>23</td>
</tr>
<tr>
<td># New sites</td>
<td>16</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>System submissions</td>
<td>212</td>
<td>8192</td>
<td>1431</td>
</tr>
</tbody>
</table>

#### 4. Metric

The overall performance measure in the second phase of the i-vector challenge was the same as in the first phase. It was based on a decision cost function (DCF) representing a linear combination of the miss and false alarm error rates at a threshold:

\[
DCF(thresh = t) = \frac{\text{misses}(thresh = t)}{\text{target trials}} \times 100 + \frac{\text{false alarms}(thresh = t)}{\text{non-target trials}}
\]

The minimum DCF (minDCF) obtained over all threshold values was the official system score recorded for a submission.

#### 5. Results

This section provides an analysis of results from the second phase of the i-vector challenge. The trials were randomly divided into two subsets, 40% in the progress set and 60% in the evaluation set. The former was used to monitor progress on the challenge scoreboard viewable by participants, while the latter was used to generate the official final scores at the end. Results on the progress and evaluation sets for the first phase can be found in [3].

![Figure 1: minDCF over time. The blue line shows the lowest minDCF on a given day on the progress set. The red line shows the minDCF for the participant with the leading performance on the progress set at the end of the evaluation period. The green line shows the lowest minDCF in a given day on the eval set.](image)

In Figure 1 we see the lowest minDCF value at any given time among all submitted systems, along with the minDCF 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 the first week and half. One notable participant registered 6 days before the end of the evaluation period and managed to end up in the top ten performers at the end of the second phase.

In Figure 2 we see the lowest minDCF at any given time over all submitted systems, as well as the minDCF at that time for this participant. The data in Figure 2 is taken over a period of 6 days. The participant submitted 12 systems in 2 days and saw a relative improvement of about 35%, from 0.349 to 0.229.

In Figure 3 we see performance on the evaluation set in the first and second phase. Of the 18 participants who submitted systems during both phases of the challenge, 15 were able to improve their system performance in the second phase.

In Figure 4 we see the improvement on the progress set generally translates into an improvement on the evaluation set. Participants 1 and 2 are from the same site; the 16 remaining are from distinct sites. Of those who participated in both the first and second phases, seven were among the top ten performers at the end of the second phase. Of the top ten
performers at the end of the first phase, only one remained in the top ten at the end of the second phase.

Figure 5 shows the minDCF of the top ten systems in phase two, and of the baseline and oracle systems. Of the 23 participants in the second phase, only one did not outperform the baseline system. The baseline system ranked 112th out of 145 systems at the end of the second with a minDCF of 0.378 on the evaluation set. It should be noted that the baseline system did not make use of the speaker labels in the development data, making its low ranking here unsurprising.

The oracle system, which did use the development data speaker labels, was outperformed by five systems on the evaluation set. The leading system at the end of the second phase had minDCFs of 0.214 and 0.195 on the progress and evaluation sets respectively. This marks a 48% relative improvement over the baseline system on the evaluation set, and a 5% relative improvement over the oracle system on both the evaluation and progress sets.

The results presented in [3] from the first phase of the i-vector challenge suggest that many systems experienced performance degradation on the evaluation set due to overfitting to same phone trials on the progress set. This trend from phase one was also observed in the second phase, as seen in Figure 4. However, it should be noted that such degradation was not as pronounced in the second phase, with the gap in performance in the different and same phone trials increasing from 0.001 to 0.041.

Systems did not demonstrate equal performance across male-only and female-only trials during the first phase, as documented in [3]. Figure 7 shows that such mis-calibration with respect to speaker sex was not as pronounced during the second phase. The leading phase two system had minDCFs on
Figure 5: The minDCF for the ten top performing in the second phase of i-vector challenge.

Figure 6: Top performing systems for first phase and second phase on the progress and evaluation trials for same and different phone numbers.

Figure 7: Leading system performance on the first and second phases on the progress set for male-only, female-only, same-sex, and all trials.

Female-only and male-only progress set trials of 0.249 and 0.250 respectively. This suggests that the labeled training data aided in the task of calibration.

6. Conclusions and Future Work

Among the goals of the first phase of the i-vector challenge was to foster research in unsupervised methods for speaker recognition. The second phase sought to determine how much better system performance could be made by exploiting speaker information. It was found that access to development speaker labels enabled a 26% relative performance improvement.

At the time of this paper’s publication, the challenge website remains operational, allowing participants to download the challenge data, submit system outputs, and view their results, and there are no plans to take the website down in the immediate future. It is our hope that leaving this resource available will assist in future exploration of ideas.

Beginning in mid-2015, NIST will launch a new i-vector machine learning challenge, focused on language recognition. Similar to the Speaker Recognition i-Vector Machine Learning Challenge, it will be meant to foster interest from the broader machine learning community, though the task in the first phase will be semi-supervised rather than supervised. For more information on the upcoming challenge, see [8].

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


