The 2013-2014 Speaker Recognition i-vector Machine Learning Challenge

1 Introduction

The National Institute of Standards and Technology (NIST) will coordinate a special i-vector challenge in late 2013 and 2014 based on data used in previous NIST Speaker Recognition Evaluations (SREs). This challenge is intended to foster interest in this field from the machine learning community and its early results will be featured in a special session of the Odyssey 2014 International Workshop on Speaker and Language Recognition.[1] It will be based on the i-vector paradigm widely used by state-of-the-art speaker recognition systems. By providing such i-vectors directly and not utilizing audio data, the evaluation is intended to be readily accessible to participants from outside the audio processing field.

The i-vectors supplied will be based on 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[2]. The mathematical basis of the i-vector system can be found here:

http://groups.csail.mit.edu/sls/publications/2011/Dehak_IEEE_May2011.pdf

Registered participants may offer multiple challenge submissions (up to 10 per day). A leaderboard will be maintained by NIST indicating the best submission performance results thus far received and processed.

2 Technical Objective

This challenge focuses on the development of new methods for using i-vectors for speaker detection in the context of conversational telephone speech. It is designed to foster research progress, including goals of:

- Exploring new ideas in machine learning for use in speaker recognition
- Making the speaker recognition field accessible to more participants from the machine learning community
- Improving the performance of speaker recognition technology.

3 Task

The challenge focuses on the task of **speaker detection**. This task is to determine whether a specified speaker (i.e., the *target speaker*) is speaking during a given segment of conversational speech.

The challenge consists of a sequence of trials. Each trial consists of a *target speaker model* (as defined by five i-vectors derived from conversations of a single target speaker) and a *test segment* defined by a single i-vector. The system must determine whether or not the speaker in the test segment is the target speaker.

For each trial the system must submit an *output*, consisting of a single (real) number, where higher numbers indicate greater degree of belief that the target speaker is the speaker in the test segment.

Performance results will be evaluated over all trials, and also over subsets of trials representing conditions of particular interest.

4 Performance Measure

The trials of the challenge will consist of a mix of target and non-target trials. *Target trials* are those in which the target and test speaker are the same person. *Non-target trials* are those in which the target and test speaker are different persons. A decision to accept or reject a trial is made by comparing a system's output to a threshold; an output greater than the threshold means accept a trial as a target trial. When a target trial is incorrectly rejected, this is a *miss* error. When a non-target trial is incorrectly accepted, this is a *false-alarm* error. By using the sorted values of outputs from a system as thresholds, the system's misses and false-alarms can be accumulated at all possible a-posteriori thresholds.

The overall performance measure will be based on a *decision cost function* (DCF) representing a linear combination of the miss and false alarm error rates at a threshold. This cost function for this challenge will be:

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

The minimum DCF obtained over all threshold values will be the official system score recorded for a submission.

Thus for each challenge participant, the performance score returned for a system submission will be this minimum DCF over the set of trials used for progress scoring. At the conclusion of the challenge, the score for a site's final submission will be determined based on the evaluation set of trials.

5 Data

The data provided will consist of development data to be used for system creation, and separate evaluation data for the challenge. The speakers involved in these two data sets will be disjoint.

The i-vectors will be derived from conversational telephone speech data in the NIST Speaker Recognitions (SRE's) from 2004 to 2012. Each i-vector will be a vector of 600 components. Along with each i-vector, a single item of metadata will be supplied, namely the amount of speech (in seconds) used to compute the i-vector. Segment durations were sampled from a log normal distribution with a mean of 39.58 seconds.

5.1 Development Data

A large quantity of additional unlabeled i-vectors from telephone call segments will be provided for general system development purposes. These will be from telephone segments of unspecified speakers and may be used, for example, for unsupervised clustering in order to learn wanted and unwanted variability in the i-vector space.

[1] See http://cs.uef.fi/odyssey2014/

[2] See http://www.nist.gov/itl/iad/mig/sre12.cfm

5.2 Evaluation Data

This will consist of sets of five i-vectors defining the target speaker models and of single i-vectors representing test segments. The number of target speaker models[3] will be 1,306 (comprising 6,530 i-vectors) and the number of test i-vectors 9,634 (one i-vector each).

The five i-vectors defining a given target speaker model will generally be chosen from conversations utilizing a common telephone handset. The handset used in the target speaker model may or may not match the handset used in a given test segment.

5.2.1 Trials for Submission and Scoring

The full set of trials for the challenge will consist of all possible pairs involving a target speaker model and a single i-vector test segment. Thus the total number of trials will be 12,582,004.

The trials will be divided into two subsets: *progress subset*, and *evaluation subset*. The progress subset will comprise 40% of the trials and will be used to monitor progress in the scoreboard. The remaining 60% of the trials will form the evaluation subset, and will be used to generate the official final scores determined at the end of the challenge.

6 BASELINE SYSTEM

A baseline system will be included in the download package. This system will serve as an example of how to achieve a successful submission. Also, the performance of the system will be the baseline for the scoreboard. The algorithm used in the baseline is a variant of cosine scoring with the following recipe:

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.

7 RULES

Each participant must complete the online registration process and download the development data, and the evaluation model and test segment i-vectors.

Each uploaded system submission must contain outputs for the full set of trials in order to be scored.

The output produced for each trial must be based solely on the training and test segment i-vectors provided for the trial (along with the development data). The i-vectors provided for other trials may not be used in any way. For example:

o Normalization over multiple test segments is not allowed.

o Normalization over multiple target speakers is not allowed.

o Use of evaluation data for impostor modeling is not allowed.

Participating sites will be limited to submitting a maximum of ten sets of system results per calendar day.

8 SYSTEM DESCRIPTION

Each participant is asked to provide an overall description of the algorithm and procedures used to create the submitted systems which may be shared with the community. Please send system descriptions to ivector_poc@nist.gov.

9 ODYSSEY 2014

Odyssey 2014: The Speaker and Language Recognition Workshop will be held June 16-19, 2014 in Joensuu, Finland. This will be the next in a series of bi-annual workshops dating back to 1994. It will feature a special session on the early results of this i-vector challenge. The site with the best performing system submitted by April 07, 2014 (by the measure specified in section 4) will receive a prize of one free Odyssey 2014 registration. The accepting site will be expected to give a presentation at the i-vector special session describing the system's algorithm and the process used to develop it. Meanwhile, regular papers for this session must be submitted for consideration by February 10, 2014[4].

10 SCHEDULE

Late-Nov-2013:	Registration opens on website
Late-Nov-2013:	Challenge data available on website
10-Feb-2014:	Odyssey papers on Challenge due
07-Apr-2014:	Last day to submit output for official scoring
08-Apr-2014:	Official scores (on evaluation subset) posted
16-Jun-2014:	Odyssey Workshop begins
30-Jun-2014:	Website closes

[3] Note that there might be multiple target models for a single person.

[4] Please note that an earlier deadline applies for submissions to other sessions of the Odyssey Workshop.