NIST Language Recognition Evaluation – Plans For 2015

Alvin F. Martin1, Craig S. Greenberg1, John M. Howard1*, Désiré Bansé1#, Lisa P. Mason2

1 The National Institute of Standards and Technology, Gaithersburg, Maryland, USA
2 U.S. Department of Defense, Fort Meade, Maryland, USA
*Contractor  #Guest Researcher
alvin.martin@nist.gov, craig.greenberg@nist.gov, john.howard@nist.gov,
desire.banse@nist.gov, lpmicke@tycho.ncsc.mil, gmason@comcast.net

Abstract

We discuss two NIST coordinated evaluations of automatic language recognition technology planned for calendar year 2015 along with possible additional plans for the future. The first is the Language Recognition i-Vector Machine Learning Challenge, largely modeled on the 2013-2014 Speaker Recognition i-Vector Machine Learning Challenge. This online challenge, emphasizing the language identification task, is particularly intended to attract interest from the machine learning community and others beyond the audio processing community. The second is the next NIST Language Recognition Evaluation, following in the series of NIST evaluations previously held in 1996, 2003, 2005, 2007, 2009, and 2011. This evaluation will emphasize language detection in the context of closely related language pairs and is open to all interested in participating.

Index Terms: i-vector challenge, language recognition evaluation, LRE

1. Introduction

Beginning in 1996 the National Institute of Standards and Technology (NIST) has coordinated six evaluations of language recognition technology. NIST specified the languages to be recognized, provided appropriate digital audio data for development and supplied the digital audio test segments to be processed. Participating sites were asked to provide decisions and scores for each segment, which were then evaluated by NIST. These results were presented and analyzed by NIST, and the systems discussed by participants, at follow-up workshops. These language recognition evaluations are referred to as LRE96, LRE03, LRE05, LRE07, LRE09, and LRE11. See [1,2,3,4,5].

Here we discuss plans for two evaluations to occur in 2015 and some plans for future evaluation. The second will be NIST’s next general language recognition evaluation (LRE15), similar in many ways to its recent predecessors. The first will be a language recognition i-vector challenge that will resemble the speaker recognition i-vector challenge conducted in 2013-2014 [6] and intended, like its predecessor, to attract interest and participation from a broader community of interest beyond that regularly involved in speech processing tasks.

In section 2 we describe further the speaker recognition i-vector challenge as background to this year’s related challenge. In section 3 we discuss how the language identification i-vector task will operate in 2015. Section 4 presents information about the web site and its leaderboard. Section 5 covers the recent evolution of the LRE’s to heavily emphasize the discrimination of closely related languages. Section 6 describes the data collected by the LDC that will be used in LRE15. Section 7 discusses the rules and procedures to be followed in LRE15. Section 8 provides some summation and indication of future language recognition plans, while section 9 provides a standard disclaimer.

2. The Speaker i-Vector Challenge

From December 2013 to June 2014 NIST coordinated Phase 1 of a special i-vector challenge [7,8] based on the i-vector paradigm widely used by state-of-the-art speaker recognition, and language recognition, systems. As with NIST’s regular series of speaker recognition evaluations (SRE’s [9]), it sought to foster research progress in speaker recognition technology; but it also sought to attract and involve the general machine learning community. To be accessible to a wider outside community it was run entirely online, and the data consisted solely of i-vectors with no audio recordings. This also implied a comparison of system outputs based on a consistent front-end and a fixed amount and type of training data.

The task was basic speaker detection. All data came from prior SRE’s. A large unlabeled development set of i-vectors was provided (disjoint from the target and test segment speakers). Each target speaker was represented by five i-vectors, each from a separate session. Test segments were each a single i-vector, and trials consisted of a combination of a target model and a segment. Along with the i-vectors, systems were provided with the segment durations, which were sampled using a log normal distribution, with a mean of approximately 40 seconds.

Forty percent of the trials were randomly assigned to a “progress set”, and the progress set performance scores were immediately displayed on the challenge leaderboard. The other sixty percent made up an “evaluation set”, and the evaluation set scores were presented at the conclusion of the challenge. Participants did not know which trials were assigned to which set.

There were 297 registrants from 47 countries, and over 8000 system submissions, numbers exceeding any of the regular NIST SRE’s. Figure 1 illustrates the tracking of best system performance (in terms of the official challenge metric) over time. The same will be done in the language challenge.
Figure 1: Tracking the min DCF (Decision Cost Function) – the primary metric for the Speaker Recognition Machine Learning Challenge. 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.

Phase 2 of the challenge began 22 July 2014 with the release of speaker labels for the development data, and continued until 15 September 2014. Systems achieved considerable further performance improvement on this now supervised learning task, See [10]. The speaker challenge website remains in place to support ongoing research efforts.

3. The 2015 Language i-Vector Task

Like the Speaker Recognition i-Vector Machine Learning Challenge, the Language i-Vector Machine Learning Challenge will utilize single i-vectors representing speech test segments (either from a telephone or from a narrowband broadcast channel) of variable lengths sampled from a log-normal distribution and drawn from conversational telephone speech selected from prior NIST evaluations. The challenge task, however, will be one of open-set language identification.

The 400-element i-vectors supplied will be produced based on a system developed by the Johns Hopkins University Human Language Technology Center of Excellence in conjunction with MIT Lincoln Laboratory [11].

Registered participants will be invited to offer large numbers of challenge submissions (The ten per day limit of the speaker challenge may be increased by as much as a factor of ten.). A leaderboard will be maintained by NIST indicating the best submission performance results thus far received and processed.

For all test segments (perhaps around 6,500) the system will be expected to output the name of the target language it is deemed to represent. Since the task is open set identification, for some segments the “out-of-set” language will be the appropriate output.

There will be 50 designated target languages, each corresponding to one of the target languages/dialects used in the NIST LRE’s between 1996 and 2011. Each target language will be defined by about three hundred i-vectors, each of these corresponding to a training segment.

In addition, participants will be provided with a development set consisting of numerous (perhaps also around 6,500) unlabeled i-vectors representing segments for general system development purposes. These may be used, for example, for unsupervised clustering in order to learn about wanted and unwanted variability in the i-vector space. This development set will include all of the target languages and multiple other “out-of-set” languages.

Participants will also be asked to provide system descriptions for their preferred systems.

4. Language i-Vector Web Site and Scoring

The primary scoring metric, to be displayed on the challenge leaderboard, will be the percentage of test segments assigned to the correct target language. Results will also be displayed for the percentage correctly classified when test segments are limited to those not corresponding to the out-of-set language (closed set scoring), and for other subsets of the test segments representing conditions of interest.

Similar to the speaker challenge, a random thirty percent of the test segments (not revealed to participants) will be assigned to a progress set used to determine the scores displayed on the leaderboard. The remaining segments will form an evaluation set for which official performance results will be announced at the conclusion of the challenge.

A baseline system will be included in the download package to serve as an example for successful submissions and as a performance comparison benchmark on the leaderboard. This will utilize a simple cosine scoring based algorithm. It is somewhat similar to the baseline system provided for the speaker challenge, and can offer reasonable performance.

Figure 2 presents a screenshot of the homepage for the challenge. It will commence in May 2015 and should conclude around September of 2015, but the web site is expected to remain available to support ongoing research effort beyond this date.

5. Emphasis on Related Languages in Recent LRE’s

The tests of the NIST LRE’s have required systems to determine for each of a collection of test segments and each of
a set of target languages (or dialects) whether or not the segment contains speech of the target, providing a hard decision and a likelihood score. In different LRE’s the tests have sometimes involved large sets of fairly distinct languages, and sometimes smaller sets or pairs of more closely related languages or dialects. The segments have involved narrowband conversational speech collected over telephone or broadcast channels. In general, there have been separate tests over segments containing approximately 3, 10, or 30 seconds of conversational speech in a channel.

The earlier LRE’s made a clear distinction between languages and dialects, and had separate tests for each. Indeed, LRE07 had separate tests of general language recognition, Chinese language recognition (Cantonese, Mandarin, Min and Wu), and dialect recognition (including Mandarin dialects, among others) [3]. But there was always some realization that this could never be an absolute and hard distinction, that issues of sovereignty and group identity play a role in the use of these terms, and that the notion of mutual intelligibility can be rather nebulous, as discussed in [5].

Further, over the course of the earlier LRE’s, system performance became increasingly strong at distinguishing languages that were not themselves fairly closely related, with performance approaching perfection for 30 s segment durations [5]. Thus, beginning with LRE09, the language/dialect distinction was abandoned, and increased emphasis was placed on distinguishing more closely related languages, whether they be viewed as truly different languages or only as different dialects.

In LRE09 there were 23 target languages, and general language recognition remained the primary task. But 16 of the 23 were selected from eight pairs of closely related languages, and participants were encouraged to do the language pairs task, if not for all 253 pairs, then for some or all of these eight pairs of particular interest, and a number of sites did this. Six of these eight proved to be among the eleven pairs most confusable pairs for one system that attempted all of the 253 pairs.

In LRE11 there were 24 target languages, and language pair distinction, across all 276 pairs, was the required task. Of the 24, 19 were from five clusters of two to five languages in each, and most of the highly confusable pairs were within a cluster. The official scoring was based on an average confusability rate across a system’s 24 most confusable pairs. See [5].

LRE15 will build on the experience of LRE09 and LRE11, with all speech selected from within clusters of related languages. The primary metric will return to one involving general language recognition over multiple languages as in LRE09, but the scoring will be limited to the measuring language distinguishability within the clusters of closely related languages.

### 6. Data for LRE15

The test data for LRE15 has been collected by the Linguistic Data Consortium (LDC), as for most of the past LRE’s. Twenty target languages, all belonging to clusters of more or less closely related languages, were carefully selected for this evaluation. There will be six such clusters involved, with two to five languages in each cluster. Table 1 provides a summary of these.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard</td>
</tr>
<tr>
<td>Chinese</td>
<td>Cantonese, Mandarin, Min, Wu</td>
</tr>
<tr>
<td>English</td>
<td>British, General American, Indian</td>
</tr>
<tr>
<td>French</td>
<td>West African, Hattian Creole</td>
</tr>
<tr>
<td>Slavic</td>
<td>Polish, Russian</td>
</tr>
<tr>
<td>Spanish (Iberian)</td>
<td>Caribbean, European, Latin American, Brazilian Portuguese</td>
</tr>
</tbody>
</table>

All of these language clusters and, with a couple of exceptions, all of these languages have been included in prior NIST LRE’s. Thus, for these exceptions will some of the newly collected data need to set aside as training data and made available for this purpose to LRE15 registered participants. The rest may be used as test data or possibly saved for future evaluation use.

The collection paradigm followed has been similar to that used in LRE11 and LRE09. A mix of conversational phone calls and narrowband broadcast segments is included for most languages. All of the data will be provided in files using NIST SPHERE format with an 8 kHz sampling rate and 16-bit PCM quantization.

Every possible effort is being made to collect unique speakers. For telephone, certain speakers (called *claque*) are paid to make multiple calls, but only their unique interlocutors are being used for test speech segments. Each call or broadcast segment is expected to provide one or more test segments of varying duration.

An extensive auditing effort assures that the segments selected contain appropriate quantities of reasonably clear speech in the specified target languages. Careful documentation of the auditing results will assist in analysis of evaluation results and future use of the data.

For each of the six clusters, it will be possible to compare performance for certain pairs with that observed on the same pair or pairs in LRE07, LRE09, or LRE11. Performance on the American/Indian English pair, in particular, was included in each of these three prior evaluations, with an observed trend of best system performance improvement over the course of these evaluations.

### 7. LRE15 Rules and Procedures

LRE15 will follow a different procedure from that of past LRE’s with respect to the training data that may be employed to develop participants’ primary systems. In the past participants were free to use all past LRE data and also welcome to use additional data as long as the data used was discussed in their system descriptions. In LRE15, primary systems will be required to use only the training data specifically provided for each target language. In most cases this will be a subset of that available from prior LRE’s and selected additional sources, and of roughly equal quantity for most of the target languages. This will remove the data acquisition aspect from the primary system competition to allow more focus on algorithm issues. Participants will be welcome, however, to submit alternate systems that use additional data, and thus be able to show what advantages may be obtained with increased data. These systems may be important for good comparisons with results from prior evaluations.
In another key change, the test segments will not be limited to speech durations of approximately 3 s, 10 s, and 30 s as was the general practice in past LRE’s. Segments will be selected to cover a broad range of speech durations including, but not limited to, those used in past practice.

LRE15 will emphasize language discrimination within each cluster, but systems will not be told the cluster to which each test segment belongs. Scoring information will therefore have to be provided for all languages for each test segment. Rather than require both a hard decision and a score for each hypothesized target language, as in past NIST LRE’s, in LRE15 only scores will be requested, and these will have to be estimated log likelihood ratios. Specifically, for each test segment each system will be asked to submit a language score vector

\[(l_1, l_2, \ldots, l_{20})\]

where \(l_i\) is its estimated log likelihood ratio for the true language of the segment being the ith target language, using the order of languages listed in Table 1.

Given such score vectors, various metrics may then be applied to measure system performance on subsets of the test segments (limited, for example, by duration) and/or subsets of the target languages (limited, for example, to a single cluster or single language pair). The primary metric for LRE15 is expected to be the closed set multi-language cost function denoted \(C_{avg}\) utilized as primary metric in LRE09 (and earlier LRE’s), limited to each language cluster considered separately. It is a simple weighted average of all target language miss rates and all target/non-target language pair false alarm rates. (See [12], section 4.1.)

But other types of scoring can be implemented as well, including using a mean of all language pair error rates (within a cluster) as done in the language pairs tests in LRE09 and LRE11. More information theoretic measures may be used as well (see [13,14]). The effectiveness of different measures may be discussed at the LRE15 workshop.

The main emphasis will be on performance measures for each cluster, but corresponding overall measures may then be defined as the means of the six cluster measures. In particular, the mean of the six single cluster primary metrics will provide a single overall system score.

In other respects, the rules for LRE15 will be similar to those of prior LRE’s. Participants will submit a primary system and, if they wish, alternate or contrastive systems. Human interaction with the audio data prior to submission will not be permitted. Scores for all target languages will be required for all evaluation test segments. (This may be relaxed with respect to requiring all target languages for re-submissions of “mothballed” systems from prior evaluations, which will be encouraged.)

All participants will be expected to send one or more knowledgeable representatives to the post-evaluation workshop to offer a presentation on their system(s) and to participate in discussions on the current state of the technology and on possible future plans.

SRE15 will take place in the autumn of 2015, with registration and distribution of training data beginning in the summer. Submissions will probably be due in October, and the evaluation workshop held in early to mid December in the southern United States. In another change from past practice, the evaluation website will generally be used for registration, data distribution, and submission of results.

8. Some Future LRE Plans

The collection of appropriate audio data in sufficient quantity is generally the limiting factor in planning major evaluations of language recognition technology. The data to be used needs to represent the languages of greatest interest, but also needs to be essentially similar across languages with respect to channel type and quality, signal format, and speaking style, while providing a broad range of speakers within each language. At the very least, there must not be differences with respect to any of these factors that correlate with language, so that, for example, channel type recognition based on different countries’ phone systems will not be confounded with language recognition.

In the earlier LRE’s the LDC collected conversational phone calls collected from a common platform based in the United States. But with the development of telecommunications in the present millennium, the cost of collecting multiple phone calls, particularly when limited to one per caller, has become quite high. The “claque” approach, where one person (the claque) calls multiple contacts in a language community has been used with some success in recent LRE’s. Future LRE’s may be expected to continue to use this approach, along with selective use of narrowband broadcast sources where multiple speakers may be obtained, as in call-in programs, and on-line sources may be explored.

The i-vector challenge provides an opportunity to reduce the collection burden. By selecting different segments from conversations and updating the i-vector selection procedure, the same underlying audio may be used in multiple such challenges. If the initial language challenge proves successful in 2015 in terms of participation and performance achieved, similar challenges may be organized afterwards, perhaps on an every other year basis.

The two recent i-vector challenges will also affect other upcoming NIST evaluations, including both SRE’s and LRE’s, in important ways. They will be increasingly web-based and utilize online resources. This will include registration procedures, the acceptance of license and other agreements, the distribution of evaluation data, and the submission of results, along with the sharing of scoring and analysis tools, and the distribution of results.

9. 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 may be 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.
10. References


