The 2015 NIST Language Recognition Evaluation Plan (LRE15)

1 INTRODUCTION

NIST has conducted a number of evaluations of automatic language recognition technology, most recently in 2009 and 2011.[1] These evaluations have been designed to foster research progress, with the goals of:

- Exploring promising new ideas in language recognition.
- Developing advanced technology incorporating these ideas.
- Measuring the performance of this technology.

The 2015 evaluation is similar in many ways to NIST's two most recent language recognition evaluations. It will involve both conversational telephone speech (CTS) data and broadcast narrowband speech (BNBS) data, and will emphasize distinguishing closely related languages.

LRE15 will, however, be different from the prior LRE's in certain key respects. The core (i.e., required) testing condition will be based on the use of only limited and specified training data to develop the models for each of the target languages . The use of unrestricted training data, including from sources other than ones provided or occurring in prior evaluation corpora, will be permitted and encouraged in an alternative test condition for comparison of algorithmic and data contributions to performance.

In addition, test segments will not be limited to segments with approximately 3 seconds, 10 seconds, or 30 seconds of speech duration. Segments will be selected to cover a broad range of speech durations, and subsequent analysis will examine the effects of segment duration on performance.

Lastly, LRE15 systems will not be asked to provide hard decisions for each target language and each test segment. Instead they will only be asked to provide for each test segment a score vector, with entries interpreted as log likelihood ratios (llr) for each of the target languages.

2 THE TASK

The 2015 NIST Language Recognition Evaluation task is language detection: Given a segment of speech and a language hypothesis (i.e., a target language of interest), the task is to decide whether that target language was in fact spoken in the given segment, based on an automated analysis of the data contained in the segment. This was the primary task in LRE09 and prior evaluations. As in the recent NIST LRE's, however, the emphasis will be on making such decisions in the context of languages that are similar to each other and frequently mutually intelligible.

2.1 TEST SEGMENTS

System performance will be evaluated by presenting the system with a series of audio test segments containing speech. For each segment, the system will consider whether each of the target languages is the language actually spoken in the test segment.

2.1.1 TARGET LANGUAGES

The twenty target languages of LRE15 will be grouped into six language clusters as described in Table 1:

Table 1: Target languages and language clusters for LRE15

Cluster	Target Languages
Arabic	Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
Chinese	Cantonese, Mandarin, Min, Wu
English	British, General American, Indian
French	West African, Haitian Creole
Slavic	Polish, Russian
Iberian	Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese

While the evaluation will focus on distinguishing languages within each cluster, the cluster to which the language of each test segment belongs will not be disclosed to systems.

2.1.2 SYSTEM OUTPUT

The output from the language recognition system for each test segment will consist of a vector of scores

$$(l_1, l_2, \ldots, l_{20})$$

corresponding to the twenty target languages in the order listed in Table 1. Further, these scores should represent estimated log likelihood, ratios, using natural (base *e*) logarithms, for the corresponding languages. In terms of the conditional probabilities for the observed data relative to the alternative target and non-target language hypotheses, the likelihood ratio *(LR)* is given by:

$$LR = \frac{prob(data \mid target\ language\ hypothesis)}{prob(data \mid non\text{-}target\ language\ hypothesis)}$$

3 TEST CONDITIONS

There will be two test conditions, a fixed training data condition and a open training data condition. All participants must offer a primary system (and optionally additional contrastive systems) using only the limited and specified training data provided for each language (i.e., the fixed training test condition). Participants may also offer one or more systems using additional training data for some or all languages, as discussed in Section 5.2 (i.e., the open training test condition).

For each speech test segment and each target language the set of non-target languages of interest will be the other languages of the cluster to which the actual language of the target segment belongs as specified in Table 1. But since the cluster of the test segment is not specified, the system must provide a 20-tuple scoring vector as indicated in Section 2 that may then be used for scoring with respect to its appropriate cluster.

3.1 DURATION

As noted in the Introduction, the approximate speech durations of the test segments will not be limited to three discrete values as in the past, but will be variable and include speech durations of up to 30 seconds as in prior evaluations, and perhaps somewhat longer

[1] These two recent evaluations are described in the following documents:

www.nist.gov/speech/tests/lre/2009/LRE05EvalPlan-v5-2.pdf

www.nist.gov/speech/tests/lre/2011/LRE07EvalPlan-v8b.pdf

as well. The non-speech portions of each segment will be included in the segment, so that each test segment will be a continuous sample of the source recording. This means, particularly in the case of conversational telephone speech, that the test segments may be significantly longer than the speech duration, depending on how much non-speech is included.

To the extent possible, the segments will be defined to begin and end at times of non-speech as determined by an automatic speech activity detection algorithm.

The actual speech duration for each test segment will not be indicated.

3.2 FORMAT

All test speech segments will be presented as a sampled data stream in 16-bit 8 kHz linear pcm format. Each segment will be stored separately in a SPHERE format file.

4 EVALUATION

Each system to be evaluated must submit a complete set of log likelihood ratio score vectors, one for each test segment. Thus each test segment will have been scored against each of the target languages.

The primary performance metric used in LRE15 will be the closed set multi-language cost function denoted C_{avg} utilized in LRE09 and earlier evaluations (see section 4.2), but applied separately to each of the six language clusters.

4.1 BASIC PERFORMANCE MEASUREMENT

As noted above, systems will provide a (20 dimensional) vector of log likelihood ratio scores for each test segment, and will not provide hard decisions as in prior LRE's. For each language L, a hard decision will be inferred from its llr score l_L according to whether or not $l_L >= 0$. (A log likelihood ratio of 0 in principle implies that with equal priors and costs it is as likely as not to be language L.) Thus if L were the true language for a test segment, this would be treated as a correct detection, while a negative value would imply a miss. If L were not the true language, the result would be a false alarm or a correct rejection, respectively.

Within each cluster, pair-wise LR performance will be computed for all target/non-target language pairs (L_T, L_N), belonging to that cluster over all test segments whose actual language is one of those within the cluster. For each such test segment and each within-cluster target/non-target language pair (L_T, L_N), the corresponding scores from the scoring vector for the test segment may be examined. Basic performance will be represented directly in terms of detection miss and false alarm probabilities. Miss probability will be computed separately for each target language, and false alarm probability will be computed separately for each target/non-target language pair. In addition, these probabilities will be combined into a single number that represents the cost performance of a system, according to an application-motivated cost model:

$$C(L_T,L_N) = C_{Miss} \cdot P_{Target} \cdot P_{Miss}(L_T) + C_{FA} \cdot (1-P_{Target}) P_{FA}(L_T,L_N)$$

where L_T and L_N are the target and non-target languages, and C_{Miss}, C_{FA} and P_{Target} are application model parameters. As in the past, the application parameters will be:

$C_{Miss} = C_{FA} = 1$, and

$P_{Target} = 0.5$

Alternatively to using 0 as an absolute decision threshold as described above, we can use a variable threshold t and similarly compute $C_t(L_T,L_N)$, following practice in prior LRE's. The minimum value thus obtained might be called min $C(L_T,L_N)$.

4.2 AVERAGE PERFORMANCE

In addition to the performance numbers computed for each target/non-target language pair, an average cost performance for the cluster will be computed:

$$C_{avg} = \frac{1}{N_L}\{[C_{Miss} * P_{Target} * \sum_{L_T} P_{Miss}(L_T)] + \frac{1}{N_L - 1}[C_{FA} * (1 - P_{Target}) * \sum_{L_T}\sum_{L_N} P_{FA}(L_T, L_N)]\}$$

where N_L is the number of languages in the cluster.

The above implicitly presumes the absolute decision threshold (t = 0 as discussed in section 4.1), but $C_{avg}(t)$ may be similarly defined for a general decision threshold t, with t fixed across all language pairs. The minimum over all t may then be designated min $C_{avg}(t)$.

The basic C_{avg} scores for each cluster will serve as the primary performance measures for a system. In addition, the average of these across the six clusters will serve as a single overall performance score measure for each system

Performance will also be examined with respect to particular subsets of interest such as with constraints on speech duration, on the type of speech recorded, using subsets of the target languages within clusters including language pairs, and using cross-cluster non-target languages.

4.3 ALTERNATIVE PERFORMANCE MEASURE

In addition to the cost metric Cavg, NIST may, at its discretion, also measure the overall calibration of the llrs in terms of predicting the posterior probabilities of each trial (which is 1 for a target trial and 0 for a non-target trial). The metric used will be the cross-entropy measure defined as[2]

$$C_{llr}^{tar}(L_T) = \frac{1}{\ln 2 \cdot |S(L_T)|} \cdot \sum_{s \in S(L_T)} \ln(1 + \exp(-l_T(s)))$$

and

$$C_{llr}^{non}(L_T, L_N) = \frac{1}{\ln 2 \cdot |S(L_N)|} \cdot \sum_{s \in S(L_N)} \ln(1 + \exp(l_T(s)))$$

where $l_T(s)$ is the log likelihood ratio for target language L_T and segment s, $S(L_T)$ is the set of test segments in language L_T and $S(L_N)$ is the set of test segments not in language L_T, and *ln* is the natural logarithm function. The average measure is:

[2] The reasons for choosing this cost function, and its possible interpretations, are described in detail in the paper "Application-independent evaluation of speaker detection" in Computer Speech & Language, volume 20, issues 2-3, April-July 2006, pages 230-275, by Niko Brummer and Johan du Preez. The function is discussed in connection with language recognition in "On Calibration of Language Recognition Scores", *Proc. 2006 IEEE Odyssey – The Speaker and Language Recognition Workshop*, by Niko Brummer and David A. van Leeuwen.

$$C_{llravg} = \frac{1}{N_L} * \sum_{L_T} [\, P_{Target} * C_{llr}^{tar}(L_T) + \sum_{L_N} P_{Non-Target} * C_{llr}^{non}(L_T, L_N)]$$

4.4 GRAPHICAL REPRESENTATION OF PERFORMANCE

In keeping with practice in previous LRE's NIST may, at its discretion, generate DET curves[3] showing system performance for particular clusters as the decision threshold is varied, or for particular language pairs within clusters.

Graphs based on the C_{llr} cost function, somewhat analogous to DET curves, may be generated, at NIST's discretion. These can serve to indicate the ranges of possible applications for which a system is or is not well calibrated.[4]

5 DATA

This evaluation will utilize both telephone bandwidth broadcast radio speech and CTS, similar to the two most recent NIST language evaluations.

5.1 LICENSE AGREEMENT

All evaluation participants, whether or not they are members of the Linguistic Data Consortium (LDC), will be required upon registrationto complete an LDC license agreement that will govern the use of all of the data supplied for use in this evaluation.

5.2 TRAINING AND DEVELOPMENT DATA

A major difference in LRE15 from prior NIST language recognition evaluations is that specific and limited amounts of training data will be provided to registered participants to develop the target language models for systems fulfilling the required fixed training data test condition (as indicated in Section 3), and such systems may not utilize any data beyond that provided. For most of the target languages this data will come from data used in prior NIST evaluations.

The training data provided may consist of extended segments, such as full phone calls in the case of CTS. Within these, the LDC will have audited for the presence of specific segments of up to 30 seconds of speech duration containing speech in the target language. Beyond these labeled segments, there could be speech present in other languages. While the training data for a language may be limited to CTS or BNBS, this does not imply that evaluation test segments will be similarly limited.

Specifics of the types and quantities of training segments to be provided for development purposes in each of the target languages will be announced subsequently. All participants registering for this evaluation and signing the LDC license agreement will automatically receive online access to this data.

Additional training data for systems for the open training data test condition may come from any source, but must be disclosed in the system description (see System Description, section 6.4, below). Its public availability or potential availability should be indicated.

5.3 EVALUATION DATA

The evaluation test segment data, collected and audited by the LDC, will be provided by NIST in the format described in section 6.2. The data will include test segments in all of the target languages included in the evaluation. The total number of evaluation test segments of all durations will not exceed 60,000. All segments will be in 16-bit 8-kHz linear pcm SPHERE format, and segments derived from CTS will not be distinguished from segments derived from BNBS.

6 PARTICIPATION INFORMATION

6.1 RULES OF PARTICIPATION

The following are rules and restrictions on system development and test, similar to those of prior evaluations. They must be observed by all participants:

- For each evaluation test segment the information available to the system is limited to that segment only (along with the training data); scores for a particular test segment must be computed without benefit from any information that might be derived from other test segments.
- Listening to the evaluation data, or any other human interaction with the data, is not allowed before all test results have been submitted.
- Results must be submitted (in the format specified in Section 6.2.1) for ***all*** test segments.
- Participants may submit results for different (e.g., "contrastive") systems. These could include "mothballed" systems used in prior language recognition evaluations. However, for each test condition there must be one (and only one) system that is designated as "primary",
- Systems for the fixed training data test condition must use only the specified training data provided to develop the models for the target languages. Sites must submit at least one system to the fixed training data test condition. Systems in the open training data test condition may use other training data as well. Any such data must be described in the system description.
- All systems submitted must be described by a system description as specified in Section 6.4.
- Each participant, whether an LDC member or not, must complete the LDC license agreement governing the use of the supplied data. (See Section 5.1.)
- Each participant must register for the evaluation before the commitment deadline, by completing the online procedure on the evaluation sign-up page:[5]

  https://lre.nist.gov/participants/sign_up

- Each participating site must send one or more representatives who have working knowledge of the evaluation system to the evaluation workshop. Representatives must give a presentation on their system(s) and participate in discussions

[3] See "The DET Curve in Assessment of Detection Task Performance" in *Proc. Eurospeech 1997*, V. 4, pp. 1895-1898, accessible online at:
http://www.nist.gov/speech/publications/index.htm

[4] See the discussion of *Applied Probability of Error (APE)* curves in the references cited in footnote 2.

[5] Please send email to LRE_poc@nist.gov if there are any questions or problems concerning this or other evaluation procedures.

of the current state of the technology and future plans. Workshop registration information will be distributed to registered evaluation participants when available. The workshop will be open only to evaluation participants and representatives of interested government and supporting agencies.

- Participants may report on their own performance in the challenge, but may not make advertising claims about winning the evaluation or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113) shall be respected[6]:

  NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.

6.2 DATA FORMAT

The data for LRE15 will be distributed to registered participants online via the evaluation website[7]. The training data and the evaluation test segments will become available on the dates indicated in the Schedule (Section 6.5). Registered participants will be informed of how to access the data.

The training data will contain a top level directory and twenty subdirectories corresponding to the target languages. Each of these twenty will contain an ASCII record format file listing the training data files for the language and a **data** subdirectory. In the former each record will contain two fields, the test segment file name and an MD5 checksum. The latter will contain the actual training data files for the language. Each training file will be a 16-bit 8-kHz pcm SPHERE format speech data file.

The evaluation data will contain a top level directory with an ASCII record format file. Each record will contain two fields, the test segment file name and an MD5 checksum. This directory will have a **data** subdirectory containing the actual evaluation test segments. Each test segment will be an 16-bit, 8-kHz, pcm, SPHERE format speech data file. The names of these files will be pseudo-random alphanumeric strings, followed by ".sph".

6.2.1 SYSTEM OUTPUT FORMAT

Sites participating in the evaluation must report test results in a single results file for each system for which results are submitted. The results files submitted to NIST must use standard ASCII record format, with one record for each test segment. Each record must contain 21 tab delimited fields consisting of the test segment file name and the 20 llr score vector entries:

1. The test segment file name, without the ".sph" extension
2. Egyptian Arabic llr
3. Iraqi Arabic llr
4. Levantine Arabic llr
5. Magrebi Arabic llr
6. Modern Standard Arabic llr
7. Cantonese llr
8. Mandarin llr
9. Min llr
10. Wu llr
11. British English llr
12. General American English llr
13. Indian English llr
14. West African French llr
15. Haitian Creole llr
16. Polish llr
17. Russian llr
18. Caribbean Spanish llr
19. European Spanish llr
20. Latin American Spanish llr
21. Brazilian Portuguese llr

6.3 SUBMISSIONS

System results for LRE15 will be submitted via the evaluation website. Specific instructions will be provided to the registered participants.

6.4 SYSTEM DESCRIPTION

Sites must provide a description for each system submitted. If multiple systems are submitted, one must be explicitly designated as the primary system and the others as contrastive systems in the system description.

The purpose of the system description is to give the other participants a good sense of your system's approaches and techniques. Please keep in mind the following guidelines when writing your system description:

Write for your audience. Remember that the reader is not **you** but other system developers who may not be familiar with your technique/algorithm. Clearly explain your method so they can understand what you did.

Be as complete possible. Note that the document should be neither pseudo-code for the inner workings of your system, nor a superficial description that leaves other system developers clueless of what you did.

Include references to item(s) referred to but not described in detail in the paper.

Avoid jargon and abbreviation without any prior context.

Sites may choose to use the 2015 InterSpeech paper submission template for their system description.

The system description should, as a minimum, include the following sections:

1. Introduction
2. System A (name of system submitted)
   2.1. System description

   [Clearly describe the methods and algorithms used in system A.]

   2.2. Training data used

   [Describe all training data used in developing system A. For the primary system this must be only the data provided to participants. For contrastive systems, note

[6] See http://www.ecfr.gov/cgi-bin/text-idx?c=ecfr&rgn=div5&view=text&n16-Apr-15ode=15:1.2.2.1.1&idno=15#15:1.2.2.1.1.0.21.14

[7] This will be https://lre.nist.gov

the source(s) of any additional data, the year published, and/or any other pertinent information.]

2.3. Processing speed

[Compute the speed of language recognition, defined as the total time duration of speech processed divided by the total CPU time across all processors required to do the processing[8]. Include the specs for the CPU and the memory used.]

3. Name of another system submitted, if any

[This section is similar to section 2 but for another system (e.g., system B). If system B is a contrastive system, note the differences from the primary system. Add new section for every system you submitted.]

4. References

[Any pertinent references]

6.5 SCHEDULE (TENTATIVE)

• July 13	LRE15 Registration opens
• July 20	Target language development data available to registered participants
• September 30	Registration for LRE15 closes
• October 19	Evaluation data (test segments) available to participants
• November 3	Evaluation submissions due at NIST by 11:59 PM, EDT
• November 17	Preliminary results and answer key released to participants
• December 8-9	Evaluation workshop held in the Ocala, FL, Southeastern United States

[8] The CPU time required to perform language recognition includes acoustical modeling, decision processing and I/O and is measured in terms of elapsed time on a single CPU, start to finish. Systems that are not completely pipelined are not penalized, however, and time intervening between separate processes need not be included in tallying elapsed time. Also excluded is time spent in system initialization (e.g., loading models into memory) and in echo cancellation (to allow the use of general purpose echo cancellation software not optimized for speed).