The 2011 NIST Language Recognition Evaluation Plan (LRE11)

1 Introduction

NIST has conducted a number of evaluations of automatic language recognition (LR) technology, most recently in 2009.[1] These evaluations are 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 2011 evaluation will differ from the prior ones in emphasizing the language pair condition, which was introduced in LRE09. Like LRE09 it will involve both conversational telephone speech (CTS) and broadcast narrow-band speech (BNBS), generally involving people telephoning into the broadcast studio. Multiple broadcast sources will be included. Unlike prior evaluations, all evaluation data will be in provided in 16-bit 8 KHz format.[2]

2 The Task

The 2011 NIST language recognition evaluation task is language detection in the context of a fixed pair of languages: Given a segment of speech and a specified language pair (i.e., two of the possible target languages of interest), the task is to decide which of these two languages is in fact spoken in the given segment, based on an automated analysis of the data contained in the segment.

2.1 Trials

System performance will be evaluated over a set of trials. Trials will consist of a test segment along with a specified target language pair. The full set of trials will consist of all combinations of an evaluation test testament and a target language pair. Thus if N is the number of target languages, each test segment will be used for N * (N-1) / 2 trials.

2.1.1 System Input

The input to the LR system for each trial will comprise:

- A segment of audio signal data containing speech, and
- The identities of the two target languages denoted ***L1*** and ***L2***.

2.1.2 System Output

The output from the LR system for each trial must include:

- The decision as to whether the language spoken in the segment is *L1* or *L2*.
- A score indicating the system's confidence in its decision, with more positive scores indicating greater confidence that the segment language is ***L1***. For each ***L1/L2*** pair, these scores must be comparable across the trials for all test segments.

Sites optionally may, and are encouraged to, choose to specify that a system's scores are to be interpreted as log likelihood ratios (using natural logarithms) for scoring purposes as discussed in section 4.3.

3 Language-Pair Test Condition

3.1 Target Languages

Table 1 lists the 24 languages to be used as target languages.[3]

Table 1: The LRE11 target languages

Arabic Iraqi	English Indian	Russian
Arabic Levantine	Farsi/Persian	Slovak
Arabic Maghrebi	Hindi	Spanish
Arabic MSA	Lao	Tamil
Bengali	Mandarin	Thai
Czech	Panjabi	Turkish
Dari	Pashto	Ukrainian
English American	Polish	Urdu

3.2 Speech Segments

3.2.1 Duration

There will be three segment duration test conditions, to test system performance on different amounts of speech:

- 3 seconds of speech, nominal. (2-4 seconds actual)
- 10 seconds of speech, nominal. (7-13 seconds actual)
- 30 seconds of speech, nominal. (25-35 seconds actual)

The actual amount of speech will vary somewhat because, 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 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 that the test segments may be significantly longer than the speech duration, particularly for CTS segments, depending on how much non-speech is included.

The nominal duration for each test segment will not be identified.

3.2.2 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. Broadcast data will in general be downsampled to 8 KHz using a low-pass filter that preserves essentially the full 0-4 KHz frequency range.

[1] These evaluations are described in the following documents:
www.nist.gov/speech/tests/lre/2003/LRE03EvalPlan-v1.pdf
www.nist.gov/speech/tests/lre/2005/LRE05EvalPlan-v5-2.pdf
www.nist.gov/speech/tests/lre/2007/LRE07EvalPlan-v8b.pdf
www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf

[2] Note that BNBS will have been originally collected in various wideband formats and converted, with downsampling, to 16-bit 8 KHz samples, while CTS will have been collected in 8-bit mu-law and converted to this format.

[3] Some of these languages may not be used if sufficient test data for them is unavailable.

4 EVALUATION

Each system to be evaluated must submit a complete set of detection results. A complete set of results comprises the detection output for testing each test segment against every target language pair. Thus the number of trials in a complete set of detection results will be T * N * (N-1) / 2, where T is the number of test segments and N is the number of target languages.

Participants may optionally submit results for multiple systems, one of which must be designated as the primary system.

4.1 LANGUAGE PAIR PERFORMANCE MEASURE

Language recognition performance will be computed for each target language pair. For each pair ***L1/L2***, the miss probabilities for ***L1*** and for ***L2*** over all segments in either language will each be determined. (Alternatively, these may be viewed as the miss and false alarm probabilities for ***L1***.)

In addition, these probabilities will be combined into a single number that represents the cost performance of a system for distinguishing the two languages, according to an application-motivated cost model:

$$\begin{aligned} C(L1, L2) &= \mathrm{C_{L1}} \cdot \mathrm{P_{L1}} \cdot P_{Miss}(L1) \\ &+ \mathrm{C_{L2}} \cdot (1 - \mathrm{P_{L1}}) P_{Miss}(L2) \end{aligned}$$

where C_{L1}, C_{L2} and P_{L1} are application model parameters. Here C_{L1} and C_{L2} may be viewed as the costs of a miss for ***L1*** and ***L2***, respectively, and P_{L1} as the prior probability for L1 with respect to this language pair. These parameters will be set to give equal cost and probability to each language:

$$C_{L1} = C_{L2} = 1, \text{ and}$$
$$P_{L1} = 0.5$$

For each system, these performance statistics will be computed separately for each of the three segment duration categories. They will be computed both based on the trial decisions (actual decision operating point) and based on the minimum possible cost obtained by varying the threshold for trial scores (minimum cost operating point). The difference between these two may be viewed as the system's calibration error cost for the language pair involved.

4.2 OVERALL PERFORMANCE MEASURE

An overall performance measure for each system will be computed as an average cost for those target language pairs presenting the greatest challenge to the system.

Let N be the number of target languages included in the evaluation. For each duration, a system's overall performance measure will be based on the N target language pairs for which the minimum cost operating points for 30-second segments, as defined in section 4.1, are greatest. For each duration, the performance measure will then be the mean of the ***actual*** decision operating point cost function values over these N pairs.

Thus calibration errors will not be considered in choosing which N cost function pairs to average, but will contribute to the system's overall performance measure.

This will be the primary overall system performance measure for LRE11.

4.3 ALTERNATIVE PERFORMANCE MEASURE

As noted in section 2.1.2 sites may specify that the scores submitted represent log likelihood ratios (*llr*'s). In terms of the conditional probabilities for the observed data of a given trial relative to the alternative language hypotheses the likelihood ratio *(LR)* is given by:

$$LR = \frac{\mathrm{prob(data \mid L1)}}{\mathrm{prob(data \mid L2)}}$$

Scores that are valid estimates of *llr*'s may be viewed as more informative and useful for a range of possible applications based on varying the parameters specified in section 4.1. A further type of scoring will be performed on such submissions. An *llr*-based performance measure, which is application independent, is defined as follows:[4]

For target language pair ***L1/L2***, let $LR(L_i,s)$ be the computed likelihood ratio for target language L_i and segment *s*. And let $S(L_i)$ denote the set of test segments in language *L*.

Then define

$$C_{llr}(L1, L2) = \frac{1}{2 * \ln 2 \cdot |\mathrm{S(L1)}|} \cdot \sum_{s \in S(L1)} \ln(1 + 1/\mathrm{LR(L1,s)}) + \frac{1}{2 * \ln 2 \cdot |S(L2)|} \cdot \sum_{s \in S(L2)} \ln(1 + LR(L2, s))$$

where *ln* is the natural logarithm function.

Likewise, a calibration independent analogue of C_{llr}, denoted C_{llr}^{min}, may be defined by a process described in the paper referenced in the preceding footnote.

An overall C_{llr} measure for each duration may then be defined as the mean of the C_{llr} values over the N pairs for which the 30-second duration C_{llr}^{min} values are greatest.

4.4 GRAPHICAL REPRESENTATION OF PERFORMANCE

In past evaluations NIST has generated DET (Detection Error Tradeoff) curves[5] based on the trial scores to show the range of possible operating points of different systems. NIST will, at its discretion, generate such curves for the language pairs and segment durations of this evaluation that appear to be informative or of particular interest. Performance may be examined with respect to factors of interest such as whether BNSB or CTS is involved. Both the minimum cost and the actual decision operating points will be noted on these curves. DETs will not be pooled across different target language pairs.

Graphs based on the C_{llr} cost function, somewhat analogous to DET curves, may also be generated, at NIST's discretion. These can serve to indicate the ranges of possible applications (as defined with respect to varying the parameters specified in section 4.1) for which a system is or is not well calibrated.[6]

[4] This 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.

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

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

5 DATA

This evaluation will utilize both telephone bandwidth broadcast radio speech and CTS, as did the 2009 evaluation. A fair degree of comparability was found in LRE09, but this will be further studied in this evaluation.

5.1 LICENSE AGREEMENT

All evaluation participants, whether or not they are members of the Linguistic Data Consortium, are required to complete the LDC license agreement that will govern the use of all of the data supplied for use in this evaluation. [7]

5.2 TRAINING AND DEVELOPMENT DATA

All data provided in connection with the previous NIST language recognition evaluations is available for training and development purposes to evaluation participants. The LDC license agreement contains a check box to request this data.

Most of the target languages to be used in LRE11 are included in the previous data. No further development data will be provided for these languages. To support algorithm development for the 2011 evaluation with respect to new target languages, all registered participants will receive from NIST a single DVD containing sample speech segments in the target languages that are new to this evaluation. (The target languages listed in Table 1 that are new include the four varietiess of Arabic plus Czech, Lao, Panjabi, Polish, and Slovak)

For each of the target languages included on this DVD, there will be 80 or more labeled segments of approximately 30 seconds speech duration each that have been audited by the LDC and found to contain narrow-band speech in the target language. These segments may be all CTS in some languages, and may be all BNBS in others. Note that this will not imply that evaluation test segments in any of these languages will be limited to a single type.

Additional training data may come from any source, but must be disclosed in the system description (see System Descriptions, below) and must either be from a publicly available source or be made publicly available shortly after the evaluation workshop.

5.3 EVALUATION DATA

The evaluation test segment data, collected and audited by the LDC, will be provided by NIST on several DVD's in the format described in section 6.2. The data will include 100 or more test segments of each of the three test durations for each 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 linear pcm 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 trial the information available to the system is limited to that specified in section 2.1.1.
- Listening to the evaluation data, or any other human nteraction 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* and for all *target language pairs* included in LRE11.
- Participants may submit results for different (e.g., "contrastive") systems. These could include "mothballed" systems used in prior language recognition evaluations. However, there must be one (and only one) system that is designated as "primary". (See section 6.3.1)
- Each participant, whether an LDC member or not, is required to 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 and signing the 2011 NIST Language Recognition registration form.[8]
- Each participating site is required to send one or more representatives who have working knowledge of the evaluation system to the evaluation workshop. Representatives will be expected to give a presentation on their system(s) and to participate in discussions 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.

6.2 DATA FORMAT

The evaluation data will be distributed on several DVD's. Each will have a top-level directory denoted, for consistency with past practice, "`lre11e`***x***", where ***x*** is a digit, and used as a unique label for the disc. The data structure is as follows:

`/lre09e`***x***`/seg.ndx` – This file contains the list of the test segments on this disc. This file is an ASCII record format file. Each record will contain just a single field, namely the test segment file name.

`/lre09e`***x***`/data/` – The **data** directory will contain the speech data 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 trial. Each record must document its decision with specification of the target language

[7] The agreement will be found at:

http://nist.gov/itl/iad/mig/upload/2011_NIST_Language_Recognition_Evaluation_Agreement_v3.pdf

[8] This form is located at: http://nist.gov/itl/iad/mig/upload/LRE11RegistrationForm.pdf. The completed form (which may be filled in online) should be returned to NIST. The FAX number is 1-301-670-0939. You may send email to LRE_poc@nist.gov if other arrangements need to be made.

pair and the test segment. Each record must contain 6 fields separated by white space and in the following order:

1. The first target language ***L1***
2. The second target language ***L2***
3. The test segment file name, without the ".sph" extension
4. The decision ("***L1***" or "***L2***")
5. The score (where the more positive the score, the more likely the language is ***L1***)

6.3 SUBMISSIONS

FTP is the preferred method for submitting the test results to NIST. Specific instructions will be provided to the registered evaluation participants.

6.3.1 SUBMISSION PACKAGING

1. Create a directory that identifies the site name and a submission identifier (e.g. ***nist1***)
2. Place the system test results file in that directory. The results file should be named according to the following convention,

   <site>_{primary, contrast1, contrast2, etc.}_(llr, not_llr}.out

   (e.g**. *nist_primary_llr.out***, ***nist_contrast1_notllr.out***)

   (Here "llr" indicates that scores may be viewed as log likelihood ratios, and "not_llr" indicates the contrary.) If you submit results for a contrastive system, you must also submit the results for the primary system. The "primary" system is the one that will be used for cross-site comparisons.
3. Compress and tar the directory (e.g. tar zcvf ***nist1.tgz nist1***)
4. FTP as anonymous to JAGUAR.NCSL.NIST.GOV. Use your e-mail address as your password
5. Change directory: cd ./lre/incoming
6. Deposit tar'd file and send email to LRE_poc@nist.gov with the following information:
   a. identity of the results file
   b. the system(s) for which results have been deposited
   c. whether or not the likelihood scores submitted may be interpreted as log likelihood ratios
   d. the system description (see section 6.3.2) of the system(s) tested, as an attachment

6.3.2 SYSTEM DESCRIPTION

Sites are to provide a description for each system submitted. If multiple systems are submitted, explicitly designate one 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 what your system is 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 as possible. However, it should neither be 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 2011 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

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

   2.2. Training data used

   [Describe all training data used in developing system A. Note the source of the data, the year published, and/or any other pertinent information.]

   2.3. Processing speed

   [Compute the speed of language recognition, defined as the total amount of speech processed divided by the total amount of CPU time required to do the processing[9]. 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]

6.4 SCHEDULE (TENTATIVE)

May 2	Development data for new LRE11 languages sent to registered participants
August 1	Registration for LRE11 closes
September 1	Evaluation data arrives at participating sites
September 15	Evaluation submissions due at NIST by 11:59 PM, EDT
October 6	Preliminary results and answer key released to participants

[9] 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).

- December 6-7 Evaluation workshop held in the Southeastern United States