The 2007 NIST Language Recognition Evaluation Plan (LRE07)

1 INTRODUCTION

NIST has conducted a number of evaluations of automatic language recognition (LR) technology, most recently in 2003 and 2005.[1] The 2007 evaluation is similar in form to these previous evaluations. The most significant differences are in the increased number of languages and dialects, the greater emphasis on a basic detection task for evaluation[2], and the increased variety of evaluation conditions.

This evaluation focuses on language and dialect detection in the context of conversational telephone speech. The evaluation is 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.

2 THE TASK

The 2007 NIST language recognition evaluation task is language detection: Given a segment of speech and a language of interest to be detected (i.e., a target language), the task is to decide whether that target language was in fact spoken in the given segment (yes or no), based on an automated analysis of the data contained in the segment.

2.1 TRIALS

System performance will be evaluated by presenting the system with a set of trials. Each test segment will be used for multiple trials, with one trial for each of the target language hypotheses that the system is being tested for.[3]

2.1.1 SYSTEM INPUT

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

- A segment of audio signal data containing speech,
- The identity of the language of interest, and
- The identities of the possible languages which might be spoken.

2.1.2 SYSTEM OUTPUT

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

- The decision as to whether the language of interest was actually spoken in the segment (yes or no).
- A score indicating the LR system's confidence in its decision, with more positive scores indicating greater confidence that the segment contains speech of the target language. These scores must be comparable across all trials in each test set.

Sites may optionally choose to specify that a system's scores may be interpreted as log likelihood ratios (using natural logarithms) for scoring purposes as discussed in section 3.3.

2.2 TARGET LANGUAGES

The number of languages to be detected has been significantly increased since the last LR evaluation in 2005. There are 26 language and dialect categories that will be used as detection targets in LRE07. These are listed in Table 1.

Table 1 A list of the target languages and dialects for LRE07.

Arabic	English	Farsi
Bengali	American	German
Chinese	Indian	Japanese
Cantonese	Hindustani	Korean
Mandarin	Hindi	Russian
Mainland	Urdu	Tamil
Taiwan	Spanish	Thai
Min	Caribbean	Vietnamese
Wu	non-Caribbean	

2.3 NON-TARGET LANGUAGES

The target languages in Table 1 will also serve as the non-target (alternative hypothesis) languages. These languages form target/non-target language pairs that are of widely varying dissimilarities, including those that are mutually unintelligible and very different as well as dialect pairs that are mutually intelligible and very similar. The discrimination task is therefore being divided into 6 different tests of varying difficulty, defined so as to probe LR capabilities and performance at both ends of this spectrum of difficulty. These 6 tests are defined in terms of the selection of target and non-target languages. Two of the tests focus on recognizing languages that are mutually unintelligible, while 4 of the tests focus on distinguishing different pairs of mutually intelligible dialects that are relatively similar to each other. The selection of languages and dialects for these tests is shown in Table 2.

Table 2 indicates a language hierarchy, with sublanguages and dialects indented below their more comprehensive language categories. For the purpose of making detection decisions, these sublanguages and dialects are included in and assumed to be part of their more comprehensive language categories. Thus, for example, in the *General LR* test, both *American* and *Indian* dialects of *English* are considered to be *English*. And all

---

[1] These evaluations are described in the following documents:
www.nist.gov/speech/tests/lang/2003/LRE03EvalPlan-v1.pdf
www.nist.gov/speech/tests/lang/2005/LRE05EvalPlan-v5-2.pdf

[2] Traditionally, recognition has been posed as an *identification* task rather than a *detection* task. Real applications are usually a hybrid of the two, with the number and selection of target and non-target languages being highly variable. NIST's emphasis on detection has been influenced by the simplicity and generality of the detection task, which gives basic detailed statistics on language recognition and which can be used to estimate performance on more complex language recognition tasks.

[3] Since the task is detection rather than identification, the segment may be judged to contain the target language for more than one target language, or for none. Decisions for the different target languages should be made separately for each trial so as to optimize the system's performance with respect to the measures specified in section 3.

sublanguages and dialects under the Chinese category are considered to be Chinese.[4]

Table 2 The six LRE07 language conditions. Target and non-target languages for each test are limited to those checked.

The Test Languages	The Tests					
	General LR	Chinese LR	Mandarin DR	English DR	Hindustani DR	Spanish DR
Arabic	✓					
Bengali	✓					
Farsi	✓					
German	✓					
Japanese	✓					
Korean	✓					
Russian	✓					
Tamil	✓					
Thai	✓					
Vietnamese	✓					
Chinese	✓					
Cantonese		✓				
Mandarin		✓				
Mainland			✓			
Taiwan			✓			
Min		✓				
Wu		✓				
English	✓					
American				✓		
Indian				✓		
Hindustani	✓					
Hindi					✓	
Urdu					✓	
Spanish	✓					
Caribbean						✓
non-Caribbean						✓

2.4 Open Set Testing

Both closed-set and open-set versions of the 6 tests will be conducted. For the closed-set tests the non-target languages will be limited to those that are checked in Table 2 for the specified test. For the open-set test the non-target languages will also include all other languages listed in Table 1 and ("unknown") languages that are not listed in Table 1. These "unknown" languages will not be disclosed to participants, and training data for them will not be made available.

2.5 Speech Segment Duration

The speech segments will be taken from telephone conversations. Each segment will be limited to one side of a conversation only. These segments will be presented as a sampled data stream in standard 8-bit 8-kHz μ-law format. Each segment will be stored separately in a SPHERE format file.

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, depending on how much non-speech is included.

Unlike previous evaluations, the nominal duration for each test segment will not be identified.

3 Evaluation

Each system to be evaluated must submit at least one complete set of detection results for at least one of the six tests. A complete set of results comprises the detection output for testing each test segment against every target language in the test. Thus the number of trials in a complete set of detection results will be N_{TS} times N_L, where N_{TS} is the number of test segments to be used in LRE07 and N_L is the number of languages in the test. Note from Table 2 that N_L is 14 for the *General LR* test, 4 for the *Chinese LR* test, and 2 for the four dialect tests.

Closed-set and open-set testing are considered to be different test conditions, and therefore a different set of results are allowed for these two test conditions for a given system for each of the 6 tests.

3.1 Basic Performance Measurement

Basic pair-wise LR performance will be computed for all target/non-target language pairs. Basic LR performance will be represented directly in terms of detection miss and false alarm probabilities. For each test, 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:

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

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. For LRE07, the application parameters will be:

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

$P_{\text{Target}} = 0.5$

[4] The categorization of Cantonese, Mandarin, Min and Wu as "sublanguages" of Chinese in the General LR test is at the least artificial and arbitrary. This grouping is harmless though, serving only to reduce the error rate and to eliminate "within-Chinese" LR evaluation.

These performance statistics will be computed separately for each of the six tests, for each of the three segment duration categories, and for the closed-set versus open-set non-target language condition.

3.2 AVERAGE PERFORMANCE

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

$$C_{avg} = \frac{1}{\mathrm{N_L}} \cdot \sum_{L_T} \left\{ \begin{array}{l} \mathrm{C_{Miss}} \cdot \mathrm{P_{Target}} \cdot P_{Miss}(L_T) \\ + \sum_{L_N} \mathrm{C_{FA}} \cdot \mathrm{P_{Non-Target}} \cdot P_{FA}(L_T, L_N) \\ + \mathrm{C_{FA}} \cdot \mathrm{P_{Out\text{-}of\text{-}Set}} \cdot P_{FA}(L_T, L_O) \end{array} \right\}$$

where

N_L is the number of languages in the (closed-set) test,

L_O is the Out-of-Set "language" (including both "unknown" languages and "known" but out-of-set languages),

$$\mathrm{P_{Out\text{-}of\text{-}Set}} = \begin{cases} 0.0 \text{ for the closed-set condition} \\ 0.2 \text{ for the open-set condition} \end{cases}$$

and

$$\mathrm{P_{Non-Target}} = \left(1 - \mathrm{P_{Target}} - \mathrm{P_{Out\text{-}of\text{-}Set}}\right) / \left(\mathrm{N_L} - 1\right)$$

This average will be computed separately for each of the three segment duration categories, and for the closed-set and open-set conditions. Thus there will be a total of six average cost performance scores for each test. These scores will serve as the primary performance measures for a system.

3.3 ALTERNATIVE PERFORMANCE MEASURE

As noted in section 2.1.2 sites may specify that the likelihood 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 target and non-target hypotheses the likelihood ratio *(LR)* is given by:

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

Scores that are estimates of *llr*'s may be viewed as more informative and useful for a range of possible applications. A further type of scoring will be performed on such submissions. An *llr*-based cost function, which is not dependent on application parameters such as those specified in section 3.1, is defined analogously to the cost function of section 3.2 as follows.

Let $LR(L_T,s)$ be the computed likelihood ratio for target language L_T and segment s. And let $S(L_T)$ denote the set of test segments in language L_T.

Then define

$$C_{llr}^{tar}(L_T) = \frac{1}{\ln 2 \cdot \left|\mathrm{S(L_T)}\right|} \cdot \sum_{s \in S(L_T)} \ln(1 + 1/\mathrm{LR(L_T, s)})$$

and

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

where *ln* is the natural logarithm function. Then the *llr* average cost measure is:[5]

$$C_{llravg} = \frac{1}{\mathrm{N_L}} \cdot \sum_{L_T} \left\{ \begin{array}{l} \mathrm{P_{Target}} \cdot C_{llr}^{tar}(L_T) \\ + \sum_{L_N} \mathrm{P_{Non-Target}} \cdot C_{llr}^{non}(L_T, L_N) \\ + \mathrm{P_{Out\text{-}of\text{-}Set}} \cdot C_{llr}^{non}(L_T, L_O) \end{array} \right\}$$

3.4 GRAPHICAL REPRESENTATION OF PERFORMANCE

In past evaluations NIST has generated DET (Detection Error Tradeoff) curves[6] based on the likelihood scores to show the range of possible operating points of different systems. NIST will, at its discretion, generate such curves for the tests of this evaluation that appear to be informative. Both the minimum cost and the actual decision operating points will be noted on these curves.

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 for which a system is or is not well calibrated.[7]

4 DATA

4.1 TRAINING AND DEVELOPMENT DATA

All data provided in connection with the previous NIST language recognition evaluations is available for training and development purposes from the Linguistic Data Consortium. To obtain this data sites, whether or not they are LDC members, must complete the required license agreement governing the use of this data. (It also governs use of the evaluation data, and thus is required of all evaluation participants.) This agreement is available on the NIST web site. (See footnote 9 in section 5.1)

Additional training data may come from any source, but must be disclosed in the system description (see System Descriptions,

---

[5] 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. 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.

[6] 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

[7] See the discussion of *Applied Probability of Error (APE)* curves in the references cited in footnote 5.

below) and must either be from a publicly available source or be made publicly available after the evaluation workshop.

Some of the languages listed in Table 2 have not been included in previous NIST evaluations and thus are not included in the CD-ROM's available from the LDC described above. All sites registering for this evaluation will also receive additional CD's, available around March 1, containing 20 conversations (or in some cases 40 conversation sides from more than 20 conversations) in each of these 7 languages.[8] Subsequently, NIST will designate two segments of each of the three durations in each of these conversation sides for development purposes.

4.2 EVALUATION DATA

Evaluation data to support the formal evaluation of the language detection algorithms will be provided by NIST on a single CD-ROM in the format described in section 5.2. The data will include 80 or more test segments of each of the three test durations for each of the 26 target languages and dialects of Table 2. Also included will be segments from languages and dialects other than those listed in Table 1, for each of the three test durations. The total number of evaluation test segments of all durations will not exceed 12,000.

5 PARTICIPATION INFORMATION

5.1 RULES OF PARTICIPATION

We summarize here the basic rules and restrictions on system development and test, most of which have been specified previously. They must be observed by all participants:

- For each LR 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 experimental interaction with the data, is not allowed before all test results have been submitted.
- For each test for which system results are submitted, they must be submitted (in the format specified in section 5.2.1) for all *target languages* included in the test.
- For each test for which system results are submitted, they must be submitted (in the format specified in section 5.2.1) for all *test segments* included in the test.
- Participants may submit results for different (e.g., "contrastive") systems. However, for each test for which results are submitted, there must be one (and only one) system that is designated as "primary". (See section 5.3.1)
- Any participant choosing to participate in a future NIST Speaker Recognition Evaluation (SRE) agrees to keep LRE and SRE data separate from each other.[9]
- Each participant, whether an LDC member or not, is required to complete the LDC license agreement and its addendum. The agreement covers the data used in previous NIST language recognition evaluations while the addendum covers this year's evaluation data and new training data.[10]
- Each participant must register for the evaluation before the commitment deadline, by completing and signing the 2007 NIST Language Recognition registration form.[11]
- 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. Registration information will be posted on the NIST Language Recognition web site, when available.

5.2 DATA FORMAT

The evaluation data will be distributed on a single DVD. There will be a top-level directory denoted, for consistency with past practice, "`lre07e1`", and used as a unique label for the disc. The data structure is as follows:

`/lre07e1/seg.ndx` – This file contains the list of the test segments to be used in all of the tests. This file is an ASCII record format file. Each record will contain just a single field, namely the test segment file name.

`/lre07e1/data/` – The **data** directory will contain all the speech data test segments. Each test segment will be an 8-bit, 8-kHz, μ-law, SPHERE format speech data file. The names of these files will be pseudo-random alphanumeric strings, followed by ".sph".

5.2.1 SYSTEM OUTPUT FORMAT

Sites participating in the evaluation must report all 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 and the test segment. Each record must contain 6 fields separated by white space and in the following order:

1. The name of the test (one of the 6 listed in Table 2: "**General_LR**", "**Chinese_LR**", "**English_DR**", "**Hindustani_DR**", "**Mandarin_DR**", or "**Spanish_DR**")
2. The target language (one of the 26 listed in Table 1)
3. The non-target language condition ("**closed-set**" or "**open-set**")
4. The test segment file name, without the ".sph" extension
5. The decision ("**T**" or "**F**")

[8] These new languages are Bengali, Russian, Thai, Cantonese, Min, Wu, and Urdu.

[9] The LDC collection protocol now includes the use of some of the same speakers for conversational data for both evaluations. Thus data from conversations collected for one evaluation should not be used in connection with the other.

[10] The agreement may be found at: http://www.nist.gov/speech/tests/lang/2007/2007_NIST_Language_Recognition_Evaluation_Agreement_Final.pdf and its addendum http://www.nist.gov/speech/tests/lang/2007/Addendum_2007_LRE.pdf

[11] This form is located at: http://ww.nist.gov/speech/tests/lang/2007/LRE07RegistrationForm.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.

6. The likelihood score (where the more positive the score, the more likely the target language)

5.3 SUBMISSIONS

FTP is the preferred method for submitting the test results to NIST.

5.3.1 SUBMISSION PACKAGING

1. Create a directory that identifies the site name and the submission number (e.g. nist1)
2. Place the system test results file in that directory. The results file should follow the convention: <site>_{primary,contrast1,contrast2,etc.}.out

   (e.g. nist_primary.out, nist_contrast1.out)

   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 ./incoming/lang
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 5.3.2) of the system(s) tested, as an attachment

5.3.2 SYSTEM DESCRIPTION

Sites are to provide a description for each system submitted. If multiple systems are submitted for a particular test set, 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 readers a good sense of what your system is about. 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.
- When possible, avoid jargon and abbreviation without any prior context.

Sites are required to use the 2006 ICSLP paper submission template[12] for their system description.

[12] www.interspeech2006.org/papers

The system description should minimally 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, where it came from, 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[13]. 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]

5.4 SCHEDULE

• March 1	Training data for the "new" languages available from the LDC
• September 1	Registration for LRE-07 closes
• October 1	Test data arrives at sites
• October 17	Submissions due to NIST by 11:59 PM, EDT
• October 26	Preliminary results and answer key released to participants
• December 11-12	Evaluation workshop in the Orlando, Florida area

[13] 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).