The 2009 NIST Language Recognition Evaluation Plan (LRE09)

1 INTRODUCTION

NIST has conducted a number of evaluations of automatic language recognition (LR) technology, most recently in 2003, 2005, and 2007.[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 2009 evaluation is similar in form to NIST's previous LR evaluations. It will, however, have an increased number of target languages. Also, much of the test data will come from Voice of America (VOA) radio broadcasts. This will be in addition to conversational telephone speech (CTS), as in previous evaluations. The test segments taken from VOA broadcasts will, however, be limited to speech assessed as being of telephone bandwidth.

2 THE TASK

The 2009 NIST language recognition evaluation task is language detection: Given a segment of speech and a language hypothesis (i.e., a target language of interest to be detected), 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.[2]

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 4.3.

3 TEST CONDITIONS

3.1 TARGET LANGUAGES

Table 1 lists the 23 languages to be used as target languages.[3]

Table 1: The LRE09 target languages

Amharic	Bosnian	Cantonese
Creole (Haitian)	Croatian	Dari
English (American)	English (Indian)	Farsi
French	Georgian	Hausa
Hindi	Korean	Mandarin
Pashto	Portuguese	Russian
Spanish	Turkish	Ukrainian
Urdu	Vietnamese	

3.2 NON-TARGET LANGUAGES

Three different alternative hypothesis test conditions will be used to specify the non-target languages:

1. For each trial, the set of non-target languages will be the set of LRE09 target languages, minus the target language. This is the "closed-set" test condition. All LRE09 participants must submit results for this test condition.
2. For each trial, the set of non-target languages will be the same as 1 above, plus other "unknown" languages whose identities will not be disclosed. This is the "open-set" test condition. This test condition is optional.
3. For each trial, the set of non-target languages will be a single language. This is the "language-pair" test condition. This test condition is optional.[4]

3.3 SPEECH SEGMENTS

3.3.1 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

[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

[2] 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 4.

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

[4] This test condition supports evaluation of dialect discrimination performance and replaces the dialect test conditions used in the LRE07 evaluation. This test condition will evaluate all possible language pairings.

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.

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

3.3.2 FORMAT

All test speech 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.

3.4 LANGUAGE-PAIR EVALUATION

Of special interest is the performance on related language pairs. Performance on the following language pairs will be of particular interest:

- Cantonese – Mandarin
- Portuguese – Spanish
- Creole – French
- Russian – Ukrainian
- Hindi – Urdu
- Farsi – Dari
- Bosnian – Croatian
- English (American) – English (Indian)

(The last five of the above language pairs are generally considered to be mutually intelligible.)

4 EVALUATION

Each system to be evaluated must submit at least one complete set of detection results for the required (closed-set) test condition. A complete set of results comprises the detection output for testing each test segment against every target language. 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 LRE09 and N_L is the number of target languages.

Optionally, detection results may also be submitted for the open-set and language-pair test conditions.

4.1 BASIC PERFORMANCE MEASUREMENT

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:

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

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 LRE09, the application parameters will be:

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

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

These performance statistics will be computed separately for each test condition and for each of the three segment duration categories.

4.2 AVERAGE PERFORMANCE

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

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

where

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

L_O is the Out-of-Set "language",

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

and

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

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

4.3 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:

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

Scores that are valid 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 performance measure, which eschews the use of specific miss and false alarm costs, is defined 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

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

and

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

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

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

4.4 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]

5 Data

This evaluation will utilize telephone bandwidth broadcast radio speech for much of its training/development and evaluation data. Such speech can be easier to collect than the CTS used exclusively in prior evaluations. A study of the effectiveness of such collection, the problems that may arise from it, and the comparative language recognition performance that may result was conducted recently by researchers at the Brno University of Technology. Their report on this effort is available and may be of interest to evaluation participants.[8]

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.[9]

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 check boxes to request the pre-2007 data (sent from the LDC) and the 2007 test data (sent from NIST).

All of the previous evaluation data consists of CTS, and most the target languages to be used in 2009 are not included in the previous data. To support algorithm development for the 2009 evaluation all registered participants will receive from the LDC two corpora of broadcast data consisting of past Voice of America broadcasts in multiple languages. These corpora, which are designated **VOA2** and **VOA3**, contain speech in most of the target languages listed in Table 1 (certain languages included as target languages in LRE07 will be exceptions) and in a variety of additional languages as well. All data will be supplied in 8-bit mu-law format. **VOA3** programs will have VOA supplied language labels, while those from **VOA2** will have possible language labels created by an automatic procedure, but these labels have will not have been audited by the LDC or NIST, and may be erroneous. Some programs may contain speech in more than one language. These are very large corpora each containing thousands of broadcasts consisting mostly of broadband (not telephone bandwidth) speech. They will be supplied on two SATA drives of one terabyte each.

For each of the 2009 target languages that are included, there will be 80 or more designated labeled segments (of approximately 30 seconds duration each) in either **VOA2** or **VOA3** that have been audited by the LDC and found to contain narrowband speech in the target language. These labeled segments may be especially useful for system training.

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 after the evaluation workshop.

5.3 Evaluation Data

Evaluation data to support the formal evaluation of the language detection algorithms, mostly collected and audited by the LDC, will be provided by NIST on a single DVD 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. Also included will be segments from languages, 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 50,000. All segments will be in 8-bit mu-law format, and segments derived from CTS will not be distinguished from segments derived from (narrowband) broadcast speech data.

[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.

[8] See "Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition: Technical Report" by Oldrich Plchot, Valiantsina Hubeika, Lukas Burget, Petr Schwarz, Pavel Matejka, and Jan "Honza" Cernocky, available at: http://www.nist.gov/speech/tests/lre/2009/radio_broadcasts.pdf

[9] The agreement will be found at: http://www.nist.gov/speech /tests/lre/2009/2009_NIST_Language_Recognition_Evaluation_Agreement_Final.pdf

6 PARTICIPATION INFORMATION

6.1 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 condition for which system results are submitted, they must be submitted (in the format specified in section 6.2.1) for all *target languages* included in the test.
- For each test condition for which system results are submitted, they must be submitted (in the format specified in section 6.2.1) 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 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 2009 NIST Language Recognition registration form.[10]
- 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 a single DVD. There will be a top-level directory denoted, for consistency with past practice, "`lre09e1`", and used as a unique label for the disc. The data structure is as follows:

`/lre09e1/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.

`/lre09e1/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".

6.2.1 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 condition ("**Closed-Set**", "**Open-Set**", or "**Language-Pair**")
2. The target language (one of the languages listed in Table 1)
3. The possible non-target languages ("**all**" for the closed-set and open-set test conditions, or the alternative language taken from Table 1 for the language-pair test condition)
4. The test segment file name, without the ".sph" extension
5. The decision ("**T**" or "**F**")
6. The score (where the more positive the score, the more likely the target language)

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 the submission number (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.}.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 ./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 for a particular test set, explicitly designate one as the primary system and the others as contrastive systems in the system description.

[10] This form is located at: www.nist.gov/speech/tests/lre/2009/LRE09RegistrationForm.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.

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.

Avoid jargon and abbreviation without any prior context.

Sites may choose to use the 2009 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[11]. 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

•	March 2	Development data (**VOA2** and **VOA3**) sent to registered participants
•	April 13	Registration for LRE09 closes
•	May 4	Evaluation data arrives at participating sites
•	May 19	Evaluation submissions due at NIST by 11:59 PM, EDT
•	June 1	Preliminary results and answer key released to participants
•	June 24-25	Evaluation workshop held in the Baltimore-Washington area

[11] 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).