The 2005 NIST Language Recognition Evaluation Plan

Introduction

NIST coordinated recent evaluations of automatic language recognition systems in 1996 and 2003. These are described in "The 1996 Language Recognition Evaluation Plan" and "The 2003 NIST Language Recognition Evaluation Plan[1]. This 2005 evaluation is similar in form to these previous ones, and is intended to establish a current baseline of performance capability for language and dialect recognition of conversational telephone speech and to stimulate further research efforts in the field.

Technical Objective

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.

The Evaluation

The task to be evaluated is the detection of a given target language or dialect. Given a test segment of speech and a target language or dialect, the system to be evaluated must determine whether or not the speech is from the target language or dialect.

The target languages and dialects will include the following:

Table 1: The Target Languages/Dialects

English (American)	**English** (Indian)	**Hindi**	**Japanese**
Korean	**Mandarin** (Mainland)	**Mandarin** (Taiwan)	**Spanish** (Mexican)
Tamil			

Language Performance Metric

The performance of a detection system is characterized by its miss and false alarm probabilities; these in turn may be conditioned by the language being detected and the actual test segment language. These probabilities will serve as the basis for evaluating system performance on the language detection task. Suppose the languages under consideration are l_1, l_2, …, l_N. In addition to the target languages listed above, we may consider "Other" as a possible language class for test data, so N may be as large as 8. Performance for detection for a target language l_i will be measured using a detection cost function, $C_{Det}(i)$, which represents the expected cost of making a detection decision:

$$C_{Det}(i) = (C_{Miss} \cdot P_{Miss(i)|Target} \cdot P_{Target}) + \sum_{j \neq i} (C_{FalseAlarm} \cdot P_{FalseAlarm(i)|Non\text{-}Target(j)} \cdot (1 - P_{Target}) / (N-1))$$

[1] These are available online at http://www.nist.gov/www.nist.gov/speech/tests/lang/1996/LRE96EvalPlan.pdf and http://www.nist.gov/speech/tests/lang/2003/LRE03EvalPlan-v1.pdf.

where C_{Miss} and $C_{FalseAlarm}$ represent the relative costs of a miss and of a false alarm, respectively (independent of language), P_{Target} is the a priori probability that a trial is a target trial (again independent of language), PMiss(i)|Target is the computed percentage of misses for target trials involving l_i, and $P_{FalseAlarm(i)|Non-Target(j)}$ is the computed percentage of false alarms for non-target trials involving detection of l_i when the test segment language is l_j. For this evaluation C_{Miss} and $C_{FalseAlarm}$ will both be 1 and P_{Target} will be 0.5. Note that for the actual test set the target probability will be considerably smaller than 1/2, but that for evaluation purposes the value of P_{Target} will always be 1/2. If the number of detected target languages under consideration is M, then we take the overall average detection cost function as:

$$C_{Det} = (\sum_i C_{Det}(i)) / M$$

Dialect Performance Metric

For English or Mandarin dialect detection, we will only define a cost function over trials involving test segments in the language of interest. Note that in each case, there are then two dialects of interest. The cost function will be taken as:

$$C_{Det} = (C_{Miss} \cdot P_{Miss|Target} \cdot P_{Target}) + (C_{FalseAlarm} \cdot P_{FalseAlarm|Non-Target} \cdot (1 - P_{Target})$$

where again $C_{Miss} = C_{FalseAlarm} = 1$ and $P_{Target} = 0.5$. Here $P_{Miss|Target}$ is the computed percentage of misses for all target trials, and $P_{FalseAlarm|Non-Target}$ is the computed percentage of false alarms for all non-target trials.

Evaluation Trials

The system under test will be tested on all test segments. For each test segment, all relevant language and dialect hypotheses will be applied in turn. Thus there will be a total of K different trials for each test segment, where K is the total number of target languages and dialects that the system is designed to detect. (In this evaluation K will be 7 or 9 or 11 depending on the dialect tests a system is undertaking. See Evaluation Rules below.)

For each trial, the system under test must provide two outputs. The first output is the actual decision ("True" or "False") of whether or not the language/dialect spoken in the test segment is the target language/dialect. The second output is a score indicating the relative likelihood that the language/dialect of the test segment is the target language/dialect.

Evaluation Conditions

Signal Conditions

The speech signal to be processed will be one side of a "4-wire" conversation and will be represented as standard 8-bit 8 kHz mu-law digital telephone data. The evaluation data will be drawn primarily (but not exclusively) from conversations recently collected by the Oregon Health & Science University (OHSU). Each test segment will be prepared by using an automatic speech activity detection algorithm to identify areas and durations of speech. The test segments will be stored in SPHERE file format, one segment per file. (Unlike in preceding evaluations, areas of silence will not be removed from the segments, but segments will be chosen to contain a specified approximate duration of actual speech.) Auxiliary information will be included in the SPHERE headers to document the source file, start time, and duration of all excerpts that were used to construct the segment.

Language Constraints

The languages/dialects represented in the test data will be drawn primarily (but not necessarily solely) from the set of target languages/dialects listed in Table 1. No additional information or constraint on language/dialect will be provided to the system under test.

Test Segment Duration

The test segments will contain three nominal durations of speech, namely 3 seconds, 10 seconds, and 30 seconds. Actual speech durations will vary but will be constrained to be within the ranges of 2-4 seconds, 7-13 seconds, and 25-35 seconds, respectively. Note that this refers to duration of actual speech contained in segments, as determined by the speech activity detection algorithm; signal durations will in general be longer due to areas of silence in the segments. Shorter speech duration test segments will be subsets of longer speech duration test segments; i.e., each 10-second test

segment will be a subset of a corresponding 30-second test segment, and each 3-second test segment will be a subset of a corresponding 10-second segment. Performance will be evaluated separately for test segments of each duration.

Speaker Sex

While side knowledge of speaker sex is inadmissible information, except as determined by an automatic algorithm, performance will be evaluated for both male and female speakers separately as well as pooled. To the extent that the available data permits, the test segments in each language will be balanced by gender.

Primary Conditions

NIST often chooses to define a subset of evaluation trials as representing the primary conditions of interest in an evaluation. For this evaluation the language recognition primary condition will consist of all trials where the test segment has a duration of (nominally) 30 seconds, is in one of the specified target languages, and is from the primary test data corpus for this evaluation (see Evaluation Data, below). For each of the three durations, performance will be noted both for all test segments, and for test segments in the target languages from the primary test data.

For English dialects, the primary condition will consist of 30 second segments in English from the primary test data. For Mandarin dialects, the primary condition will consist of 30 second segments in Mandarin from the primary test data.

Corpus Support

Training Data

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. In addition, 20 (or 40) complete conversations for each of the target languages listed in table 1 are available from the LDC for research purposes.[2] These are from the CallFriend Corpus and were also made available for the previous evaluations.

Development Data

Development data, to support development, refinement, and pre-evaluation testing of language detection algorithms, will be provided by NIST on three CD-ROM's. Each CD will contain test segments taken from each of 20 conversations for each of 12 target languages.[3] Two test segments of each of the three test durations will be supplied for each side of each conversation. One of these CD's contains the development data from the 1996 evaluation; the second contains the evaluation data from the 1996 evaluation, and the third contains evaluation data from the 2003 evaluation.

Evaluation Data

Evaluation data to support the formal evaluation of the language detection algorithms will be provided by NIST on a single CD-ROM. These data will comprise 160 or more test segments of each of the three test durations for each of the 7 target languages. (For English and Mandarin, with two dialects to be tested, there may be as many as twice this number.) This primary test data will be supplemented with 640 or more segments from other languages and/or other corpora, for each of the three test durations. The total number of evaluation test segments of all durations will not exceed 12,000.

Evaluation Rules

The evaluation may be viewed as consisting of 11 different tests, one for each of the 7 specified target languages and one for each of the 4 specified dialects of English or Mandarin. Participating sites are expected to do all of the 7

[2] For three of the languages, namely English, Mandarin, and Spanish, 40 complete conversations are available, while 20 conversations are provided for the other four languages plus five additional languages. This reflects the languages and dialects included in the 1996 evaluation.

[3] For two of the CD's and three of the languages, namely English, Mandarin, and Spanish, segments are included from 40 conversations. See Footnote 1. The third CD contains some additional English and Japanese test segments.

language tests, but may choose to do, or not, the English dialect tests and the Mandarin dialect tests. They must indicate on the registration form (see Schedule below) which dialect tests they will do. For each test performed, it is imperative that system results for all (of the up to 3,120) test segments be submitted in order for that test to be considered valid and to be accepted.

The following evaluation rules and restrictions on system development and test must be observed by all participants:

- Each test segment is to be processed separately, independently, and without use of any knowledge of other test segments. Especially, normalization over multiple test segments is not allowed.
- Use of the knowledge of the whole set of target languages and dialects is allowed. Thus, normalization over multiple target languages is allowed, as is limiting (to say, one) the number of languages for which a "true" decision is made on any given test segment. Note, however that there may be test segments from nontarget languages, which are unknown to the system. Use of the knowledge of these languages is not allowed.
- Side knowledge of the sex or other characteristics of the test speaker (except as obtained by automatic means) is not allowed.
- Listening to the evaluation data, or any other experimental interaction with the data, is not allowed before all test results have been submitted.

Data Set Organization

Each of the three development data set CD-ROM's and the single evaluation data set CD-ROM's will have the same organization. Each disc's directory structure will organize the data according to information that is admissible to the language recognition system. The directory structure will be as follows:

- There will be a single top-level directory on each disc, used as a unique label for the disc. These directories will be named "**lid96d1**", "**lid96e1**", and "**lid03e1**" for the development data CD-ROM's (the names used in the '96 and '03 evaluations) and "**lre05e1**" for the evaluation data CD-ROM.
  - Under the top-level directory there will be a subdirectory named "**test**" for storing the test data.
    - Under the **test** directory there will be three duration subdirectories, namely "**30**" (for the 30 second test segments), "**10**" (for the 10 second test segments), and "**3**" (for the 3 second test segments).
      - In each of the **30**, **10**, and **3** segment duration directories will be stored the test segments. Each test segment will be stored in a SPHERE-format mu-law speech data file. The names of these files will be pseudo-random alphanumeric strings, followed by "**.sph**".

For the three development data discs only, each of the three test segment duration subdirectories will contain an index file for associating each test segment with the language spoken in that segment. This file will be named "**seg_lang.ndx**" and will use standard ASCII record format. Each record in this file will contain the name of a test segment file followed by the name of the language spoken in that file. The languages will be represented by the following character strings (these include languages not included as targets in 2005):

- "**Arabic**", "**English**", "**Farsi**", "**French**",
- **"German**", "**Hindi**", "**Japanese**", "**Korean**",
- "**Mandarin**", "**Spanish**", "**Tamil**", "**Vietnamese**"

For the evaluation data set only, each of the three test segment duration subdirectories will contain an index file, which specifies the test segments to be processed. This file will be named "**seg.ndx**" and will use standard ASCII record format. Each record in this file will contain the name of a test segment file (in the corresponding test segment directory) to be processed. The evaluation test will be to process each of the test segments named in the index file against each target language and each chosen dialect.

Format for Submission of Results

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 decision. Each record must document its decision with identification of the target language/dialect and the test segment. Each record must contain 5 fields separated by white space and in the following order:

The target language/dialect (one of "**English**", "**Hindi**", "**Japanese**", "**Korean**", "**Mandarin**", "**Spanish**", "**Tamil**", "**English.American**", "**English.Indian**", "**Mandarin.Mainland**", or "**Mandarin.Taiwan**")

The test segment duration (one of "**3**", "**10**", or "**30**")

The test segment file name, without the ".sph" extension

The decision (one of "**T**" or "**F**")

The score (where the more positive the score, the more likely the target speaker)

Results files may be submitted via FTP:

1. FTP as anonymous to JAGUAR.NCSL.NIST.GOV, use your e-mail address as your password
2. Change directory: cd ./incoming/lang
3. Deposit your results, send email to audrey.le@nist.gov, identifying your results file.

Alternatively, result files may be emailed directly to audrey.le@nist.gov

System Descriptions

Sites are to provide a brief system description for each set of results submitted. If more than one set of results is submitted, a "primary" system must be identified. The "primary" system is the one that will be used for cross-site comparisons.

Along with a brief description of the algorithmic approach and any training data utilized, sites must report CPU execution time for generating likelihood scores for the test data, as if the test were run on one CPU. Sites must also report the specs for the CPU as well as the memory, using a reporting format specified by NIST at the time of the evaluation.

Schedule

- The development data set CD-ROM is available from NIST.
- October 7th 2005:
  - Commitment deadline. All participating sites must register with NIST, by completing and signing the 2005 NIST Language Recognition registration form[4].
- October 24th 2005:
  - The evaluation data CD-ROM will be shipped via Federal Express, scheduled to arrive at the registered sites on this date. On this day the evaluation testing period will commence.
- November 7th 2005:
  - Evaluation results must be submitted to NIST no later than 12 noon, EST.
- November 11th 2005:
  - Preliminary results released to the participants. NIST will also release the answer key to facilitate diagnostic analysis of the results.

[4] This form is located at: http://ww.nist.gov/speech/tests/lang/2005/LRE05RegistrationForm.pdf. The completed form (which may be filled in online) should be returned to Alvin Martin at NIST. His FAX number is 1-301-670-0939. You may send email to alvin.martin@nist.gov if other arrangements need to be made.

- December 6th-7th 2005:
    - The follow-up workshop will be held in the Baltimore-Washington area at a location to be announced. All participants are required to send a representative who has working knowledge of the evaluation system. Registration information will be posted on the NIST Language Recognition web site, when available.