The 2003 NIST Language Recognition Evaluation Plan

Introduction

The most recent NIST Language Recognition Evaluation was held in 1996, and is described in "The 1996 Language Recognition Evaluation Plan". This new evaluation for 2003 is very similar to it, and is intended to establish a new baseline of current performance capability for language recognition of conversational telephone speech and to lay the groundwork for further research efforts in the field.

Technical Objective

The task is to detect the presence of a hypothesized target language, given a segment of conversational speech over the telephone. The target language will be one of the following twelve languages:

Table 1: The Target Languages

English (American)	**Arabic** (Conversational Egyptian)	**Farsi**	**French** (Canadian French)
Mandarin	**German**	**Hindi**	**Japanese**
Spanish (Latin American)	**Korean**	**Tamil**	**Vietnamese**

A secondary research objective is to achieve uniform performance across all target languages.

The Evaluation

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

The performance of a detection system is characterized by its miss and false alarm probabilities; thus these probabilities will serve as the basis for evaluating system performance on the language detection task. Performance will be measured using a detection cost function, C_{Det}, which represents the expected cost of making a detection decision:

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

where C_{Miss} and $C_{FalseAlarm}$ represent the relative costs of a miss and a false alarm, respectively. For this evaluation C_{Miss} and $C_{FalseAlarm}$ will both be 1 and the a priori probability of the target language will be 0.5.

The system under test will be tested on all test segments. For each test segment, all of that system's target language hypotheses will be applied in turn. Thus there will be a total of N different trials for each test segment, where N is the number of target languages that the system is capable of detecting. (Note that for the actual test set the target language probability will be considerably smaller than 1/2 , but that for evaluation the value of P_{Target} will be 1/2.)

For each trial, the system under test must provide two outputs. The first output is the actual decision ("true" or "false") regarding whether or not the language spoken during the test segment is the target language. The second output is a score indicating how likely the language of the test segment is the same as the target language.

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 necessarily exclusively) from unexposed conversations collected for the LDC's CallFriend corpus. Each test segment will be prepared by using an automatic speech activity detection algorithm to identify areas of speech. Continuous excerpts will be concatenated to produce all test segments, which will be stored in SPHERE file format, one segment per file. 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 represented in the test data will be drawn primarily (but not necessarily solely) from the set of target languages listed in Table 1. No additional information or constraint on language will be provided to the system under test.

Test Segment Duration

The test segments will be of three nominal durations, namely 3 seconds, 10 seconds, and 30 seconds. Actual durations will vary but will be constrained to be within the ranges of 2-4 seconds, 7-13 seconds, and 25-35 seconds, respectively. Shorter test segments will be subsets of longer 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.

Speaker Sex

While side knowledge of speaker sex is inadmissible information, 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.

Corpus Support

Training Data

Training data may come from any source. In addition, 20 complete conversations for each of the 12 target languages listed in table 1 are available from the LDC for research purposes.[1] These are from the CallFriend Corpus and were also made available for the 1996 evaluation.

Development Data

Development data, to support development, refinement, and pre-evaluation testing of language detection algorithms, will be provided by NIST on two CD-ROM's. Each CD will contain test segments taken from each of 20 conversations for each of the 12 target languages.[2] 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 other contains the evaluation data from this evaluation. Thus there will be a total of 7,200 development test segments (for a total of about 30 hours of speech).

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 80 test segments of each of the three test durations, for each of the 12 target languages. These primary test data may be supplemented with up to 320 segments from other languages and/or other

[1] For three of the languages, namely English, Mandarin, and Spanish, 40 complete conversations are available. This is because the 1996 evaluation included two different dialects of each of these languages.

[2] For three of the languages, namely English, Mandarin, and Spanish, segments are included from 40 conversations. See Footnote 1.

corpora, for each of the three test durations. Thus there will be a total of up to 3,840 evaluation test segments (for a total of about 17 hours of speech) and a maximum of 46,080 detection trials.

Evaluation Rules

A total of 12 tests constitute the evaluation. These tests are namely a test for each of the 12 languages. Funded contractors wishing to do fewer than the full 12 tests must get sponsor approval for the subset of tests to be conducted. For each test performed, it is imperative that system results for all 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 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 two 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**" and "**lid96e1**" for the development data CD-ROM's (the names used in the '96 evaluation) and "**lid03e1**" 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 two 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:

- "**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 a chosen target language.

Format for Submission of Results

Sites participating in the evaluation must report all test results for each test submitted. The results must be provided to NIST in results files using standard ASCII record format, with one record for each decision. Each record must document its decision with identification of the target language and the test segment. Each record must contain 5 fields separated by white space and in the following order:

The target language (one of "**Arabic**", "**English**", "**Farsi**", "**French**", "**German**", "**Hindi**", "**Japanese**", "**Korean**", "**Mandarin**", "**Spanish**", "**Tamil**", or "**Vietnamese**")

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 mark.przybocki@nist.gov, identifying your results file.

Alternatively, result files may be emailed directly to mark.przybocki@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 are 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, 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[3].
- February 19th 2003:
  - Commitment deadline. All participating sites must register with NIST, by completing and signing the 2003 NIST Language Recognition registration form[4].
- February 24th 2003:
  - The evaluation data CD-ROM will be shipped by the way of Federal Express, scheduled to arrive at the registered sites on February 24th, 2003. On this day the evaluation testing period will commence.
- March 10th 2003:
  - Evaluation results must be submitted to NIST no later than 12 noon, EST.
- March 12th 2003:
  - Preliminary results released to the participants. NIST will also release the answer key to facilitate diagnostic analysis of the results.
- **April 28th-29th 2003**:
  - The follow-up workshop will be held at NIST. 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.

[3] To acquire the development data and/or the evaluation data, participants must either be members of the Linguistic Data Consortium or sign the limited use license agreement located on the LDC website.

[4] This form is located at: http://ww.nist.gov/speech/tests/lang/doc/RegistrationForm.pdf. The completed form 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.