The 1996 Language Recognition Evaluation Plan

Introduction

In the past, Language Recognition R&D has focused on specific language identification tasks, with significant application-specific and language-specific research investment required to provide technology for each task. This year, the emphasis will be on research directed toward a general base of technology that may be ported to various language recognition tasks with minimum effort, and to develop the ability to make more difficult discriminations between similar languages and dialects of the same language. This focus augments the traditional evaluation goals, those being:

1. to drive the technology forward,
2. to measure the state-of-the-art, and
3. to find the most promising algorithmic approaches.

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 set of fifteen languages:

Table 1: The Target Languages

English (General American)	**English** (Southern American)	**Arabic** (Conversational Egyptian)	**Farsi**	**French** (Canadian French)
Mandarin (from Mainland China)	**Mandarin** (from Taiwan)	**German**	**Hindi**	**Japanese**
Spanish (Caribbean)	**Spanish** (Highland)	**Korean**	**Tamil**	**Vietnamese**

Note that this set of languages includes two dialects each of English, Mandarin, and Spanish. These dialects will be treated as separate languages, but attention will also be given to the ability to discriminate between different dialects of the same language.

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, a target language will be assigned as a test hypothesis, and the task is to determine whether this test hypothesis is true or false.

The performance of a detection system is characterized by its miss and false alarm probabilities, and these probabilities will therefore 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 expect 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 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 and dialects that the system is capable of detecting. (Note that for the actual test the target language probability will be 1/15, 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 simply a decision regarding whether 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 conversations will be drawn primarily (but not necessarily exclusively) from LDC's *CallFriend* corpus. Each test segment will be prepared by using an automatic speech activity detection algorithm to select continuous excerpts of speech. These speech regions will be concatenated to produce each test segment. The test segments so produced 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 which were used to construct the segment.

Language Constraints

The languages will be drawn primarily (but not 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. Evaluation will, however, contrast target language detection performance for various language pairs.

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.

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.

Corpus Support

Training Data

Training data may come from any source. In addition, 20 complete conversations for each of the 15 target languages listed in table 1 are available from the LDC for research purposes.

Development Data

Development data, to support development, refinement, and pre-evaluation testing of language detection algorithms, will be provided by NIST on a single CD-ROM. The test segments to be supplied will be taken from each of 20 conversations for each of the 15 target languages and dialects. Two test segments of each of the three test durations will be supplied for each side of each conversation. Thus there will be a total of 3,600 development test segments (for a total of about 15 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 15 target languages and dialects. These primary test data will be supplemented with up to 320 segments from other languages and conditions, for each of the three test durations. Thus there will be a total of up to 4,560 evaluation test segments (for a total of 18 hours of speech) and a maximum of 68,400 detection trials.

Evaluation Rules

A total of 15 tests constitute the evaluation. These tests are namely a test for each of the 15 languages and dialects. Funded contractors wishing to do fewer than the full 15 tests must get sponsor approval for the subset of tests to be conducted. For each test performed, it is imperative that ***all*** 4,560 test results 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. Note, however that there will 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 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

Both the development data set and the evaluation data set CD-ROM's will have the same organization. Each disk'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 disk, used as a unique label for the disk. This directory will be named "**lid96d1**" for the development data CD-ROM and "**lid96e1**" 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 the these files will be pseudo-random alphanumeric strings, followed by "**.wav**".

For the development data set 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.Egyptian**" "**English.Am.Gen**" "**English.Am.South**"
"**Farsi**" "**French.Canada**" "**German**"
"**Hindi**" "**Japanese**" "**Korean**"
"**Mandarin.North**" "**Mandarin.Taiwan**" "**Spanish.Highland**"
"**Spanish.Carib**" "**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. These 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:

1. The target language
2. The test segment duration (one of "**3**", "**10**", or "**30**")
3. The test segment file name
4. The decision (one of "**T**" or "**F**")
5. The score (where the more positive the score, the more likely the target speaker).

Execution Time

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 will be distributed by NIST on 23 April 1996.
- The evaluation data set CD-ROM will be distributed by NIST on 20 May 1996, at which point evaluation testing may commence.
- Evaluation results must be submitted to NIST no later than 3 June 1996. Once each site has completed its submissions to NIST, NIST will distribute an answer key to that site which identifies the language being spoken for each test segment, to facilitate diagnostic analysis of the results.
- The follow-up workshop will be held on 17-18 June 1996 at the Maritime Institute.