NIST 2007 Language Recognition Evaluation

Alvin F. Martin, Audrey N. Le

Speech Group, Information Access Division, Information Technology Laboratory
National Institute of Standards and Technology, USA
{alvin.martin,audrey.le}@nist.gov

Abstract

This paper discusses NIST's 2007 evaluation of language recognition. Some history of earlier NIST language evaluations is covered, and the test procedures and protocols, evaluation data used, and planned measures of performance for the 2007 evaluation are described. The participants and submissions of the 2007 evaluation are described, and preliminary information is included on the evaluation performance results after brief initial analysis.

1. Introduction

NIST, the National Institute of Standards and Technology, located in Gaithersburg, Maryland, USA, has coordinated previous evaluations of automatic language recognition systems in 1996, 2003, and 2005 [1]-[5]. Its fourth such evaluation is occurring in 2007 [6].

The 2007 evaluation involves 26 different target languages and dialects and six different tests of language or dialect recognition, as specified in Table 1. The general language test involves 14 target languages, including Chinese viewed as a single language class. The Chinese language test involves four target Chinese languages. For English, Mandarin (a Chinese language), Hindustani, and Spanish a dialect test involving two of each language's dialects is included.

Section 2 discusses the test protocols in greater detail.

The evaluation data consists of separate series of test segments containing approximately 30 seconds, 10 seconds, or 3 seconds of conversational telephone speech. In addition to test segments in the 26 target languages and dialects, some test segments are included consisting of speech in other "unknown" languages or dialects not specified in advance to evaluation participants.

For each segment and each target language systems must determine, via a hard decision and a score, whether or not the test segment contains speech of the target language. Trials of the three durations are scored separately, resulting in 18 different test conditions (6 tests x 3 durations).

Section 3 discusses the data in greater detail.

System performance is measured by a specified cost function that involves pair-wise language specific miss and false alarm rates and weightings across languages. An alternative measure based on likelihood ratios is also offered, and appropriate DET (Detection Error Tradeoff) curves will also be generated.

The performance measures are discussed in greater detail in Section 4.

Table 1: The six LRE07 language conditions. Target and non-target languages for each test are limited to those checked.

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

Section 5 discusses the participating sites in the 2007 evaluation and the numbers of system submission results received.

Section 6 provides an initial look at the 2007 evaluation performance results, and compares them with those of previous evaluations.

Finally, section 7 briefly addresses future evaluation plans.

2. Test Protocols

The basic task of the 2007 evaluation is language (or dialect) detection: Given a segment of speech and a language of interest (target language), determine whether or not that language is spoken in the segment, based on an automated analysis of the data contained in the segment.

System performance is evaluated by presenting the system with a sequence of trials. Each test segment is used for multiple trials, with one trial for each of the target language hypotheses being tested for.

The system input for each trial comprises:

- A segment of audio signal data containing speech
- The identity of the target language/dialect of interest for the trial
- The identities of the possible languages/dialects being spoken included in the specific recognition test involved

The required output for each trial includes:

- The decision of whether or not the language/dialect of interest is spoken in the segment (yes or no)
- A score indicating the system's confidence in the decision, with higher scores denoting greater confidence that the segment contains speech of the target language/dialect. These scores must be comparable across all trials for each test

Participants may optionally choose to specify that their systems' scores may be interpreted as log likelihood ratios (using natural logarithms) for scoring purposes as discussed in Section 4.

As noted, Tables 1 specifies the 26 target languages and dialects of the evaluation and the six different types of language and dialect tests included in the evaluation. Each evaluation system may produce output for some or all of the six tests, but must provide output for all of the trials included in each chosen test.

The inclusion of 14 languages for general language recognition represents an expansion from the 7 to 12 languages included in the previous evaluations. The inclusion of multiple Chinese languages (which are generally not mutually intelligible) is a new feature of this evaluation, as is a Hindustani dialect test involving Hindi and Urdu.

Previous NIST language recognition evaluations have concentrated on closed set testing, where trials are limited to segments whose speech is in fact in one the target languages or dialects. The 2007 evaluation permitted sites to indicate whether systems were designed primarily for the closed set condition or for open set testing, or to submit separate results for each. The test data included a large number of segments in several different languages not among those specified as target languages. These "unknown" languages were not disclosed to participants, and no training data for them was made available.

Participating sites could submit multiple systems for each test, but were required to specify a single system as primary for each condition (including test, duration, and closed vs. open).

3. Data

The speech segments are all taken from conversational telephone data. Each segment is limited to one side of a conversation only. Each is presented as a sampled data stream in standard 8-bit 8-kHz u-law format stored separately in a SPHERE format file.

As noted, there are three segment duration test conditions, designed 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 amounts of speech vary somewhat because, to the extent possible, the segments are chosen to begin and end at times of non-speech as determined by an automatic speech activity detection algorithm. Non-speech portions are included in the test segments, making each segment a continuous sample of the source recording. Thus 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 is not identified. But performance is measured separately for each of the three durations.

Table 2: Numbers of test segments (of each duration) for the target and "unknown" languages and dialects of the 2007 evaluation

Language/Dialect	Segments per duration
Arabic	80
Bengali	80
Farsi	80
German	82
Japanese	80
Korean	80
Russian	161
Tamil	168
Thai	82
Vietnamese	160
Chinese/Cantonese	80
Chinese/Min	80
Chinese/Wu	80
Chinese/Mandarin/Mainland	80
Chinese/Mandarin/Taiwan	80
English/American	80
English/Indian	160
Hindustani/Hindi	160
Hindustani/Urdu	80
Spanish/Caribbean	80
Spanish/non-Caribbean	160
French – unknown	80
Italian – unknown	80
Indonesian – unknown	80
Punjabi – unknown	32
Tagalog – unknown	80

The data used in the previous NIST language evaluations was made available to all participants, as were some conversations in target languages or dialects not included in the previous evaluations, for use in system training and development.

Most of the data used in the 2007 evaluation was recently collected by the Linguistic Data Consortium (LDC). In addition, some unreleased data previously collected by the Oregon Health Sciences University (OHSU) was also used. Canadian French data of the previously collected LDC CallFriend Corpus was used as one of the unknown languages. Table 2 lists the numbers of test segments of each of the three durations of each language or dialect included in the 2007 evaluation data. In general, about 80 segments of each duration were selected from twenty conversations (two per conversation side) of each language or dialect, with twice as many segments included when both LDC and OHSU data was available. Note that five unknown languages were included, including a relatively small amount of Punjabi speech.

4. Measures of performance

Language recognition is a detection task for which there are two types of errors. Errors in *target* trials, those for which the correct is answer is 'yes' (the target language is present) are *misses*; errors in *non-target* (*impostor*) trials, those for which the correct answer is 'no' are *false alarms*. Thus for any test condition there is a *miss rate* and a *false alarm rate*. Error cost functions are then defined as appropriate combinations of these basic error rates.

4.1 Pair-wise Measure

Basic pair-wise recognition performance is computed for all target/non-target language pairs. This represents all trials where the target language is a specified language L_T and the segment language is either L_T or a specified other language L_N. The miss and false alarm probabilities for these trials are combined into a single number that represents the cost performance for the 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 C_{Miss}, C_{FA} and P_{Target} are application model parameters. For LRE07, as in previous evaluations, these application parameters are:

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

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

Such performance statistics are computed separately for each of the six tests, for each of the three segment duration categories, and for both the closed-set and open-set non-target language conditions.

4.2 Average Performance

An average cost performance across languages is then 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" (including both "unknown" languages and "known" but out-of-set languages),

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

and

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

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

4.3 Alternative Performance Measure

As noted in Section 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:

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

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 is available on such submissions. An average *llr*-based cost function, not dependent on application parameters such as those specified in Section 4.1, is defined analogously to the cost function of section 4.2 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

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

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:

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

This reasons for choosing this type of cost function, and its possible interpretations, are described in detail in [7]. Its use specifically in connection with language recognition is further discussed in [8]. It may be noted that it is an unbounded non-negative cost function with lower values representing better performance. Its units may be viewed as bits of information, with a value of $\log_2(\mathrm{L_N})$ corresponding to a system relying on the prior alone.

4.4 Graphical Performance Representation

As in past evaluations, NIST is generating DET (Detection Error Tradeoff) curves [9] based on the likelihood scores to show the range of possible operating points of the different systems under particular test conditions. Such plots are also used to compare the performance results of the best systems in this evaluation with those of best systems in previous evaluations under similar test conditions. See section 6 below.

Graphs based on the C_{llr} cost function, known as APE (Applied Probability of Error) curves and somewhat analogous to DET curves, may also be generated. These can serve to indicate the ranges of possible applications for which a system is or is not well calibrated.

5. Participants

There were 21 participating sites in the 2007 evaluation, compared with 12 in the 2005 evaluation. These included four sites each from China, France, and Spain, three from the United States, two from Singapore, and one each from the Czech Republic, Italy, Germany, and the Netherlands/South Africa (working as a team).

6. Performance Results

Table 3 lists the best $\mathrm{C_{avg}}$ values achieved by primary systems for each of the six tests and each of the three durations for closed set testing, and Table 4 lists the corresponding best $\mathrm{C_{avg}}$ values for open set testing.

Table 3: Best primary system $\mathrm{C_{avg}}$ values for the 2007 closed set tests, with comparative results for 2005 shown in gray

	Duration		
	30 sec.	**10 sec.**	**3 sec.**
General LR	0.0103	0.0363	0.1335
2005	0.0419	0.0715	0.1569
Chinese LR	0.0490	0.0951	0.2096
English DR	0.0875	0.1250	0.2031
2005	0.0594	0.1234	0.2442
Hindustani DR	0.3156	0.3484	0.3781
Mandarin DR	0.1135	0.1788	0.2672
2005	0.1987	0.2442	0.3212
Spanish DR	0.3438	0.4000	0.4344

Table 4: Best primary system $\mathrm{C_{avg}}$ values for the 2007 open set tests

	Duration		
	30 sec.	**10 sec.**	**3 sec.**
General LR	0.0305	0.0596	0.1532
Chinese LR	0.0481	0.0802	0.1829
English DR	0.1312	0.1722	0.2373
Hindustani DR	0.2912	0.2990	0.3637
Mandarin DR	0.1217	0.1982	0.3003
Spanish DR	0.3028	0.3350	0.3761

Figures 1, 2, and 3 present DET plots of closed set general language recognition results for the best performing primary systems in the 2003, 2005, and 2007 evaluations for 30, 10, and 3 second segments, respectively. Note that the languages included vary with the evaluation year. (The numbers of target languages were 12 in 2003, 7 in 2005, and 14 in 2007.)

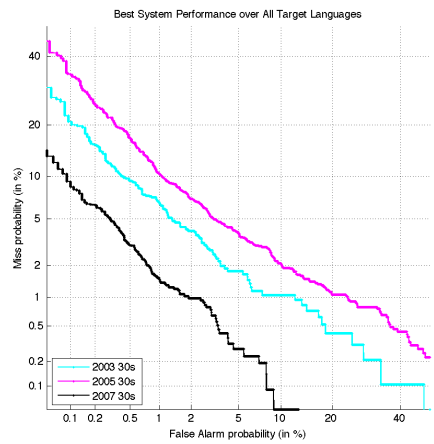

Figure 1: Best system performance on 30 second closed set general language recognition trials in the 2003, 2005, and 2007 evaluations

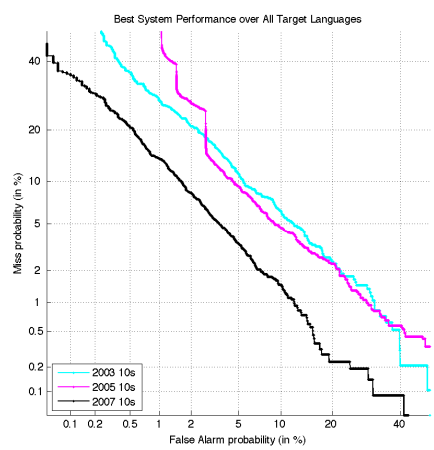

Figure 2: Best system performance on 10 second closed set general language recognition trials in the 2003, 2005, and 2007 evaluations

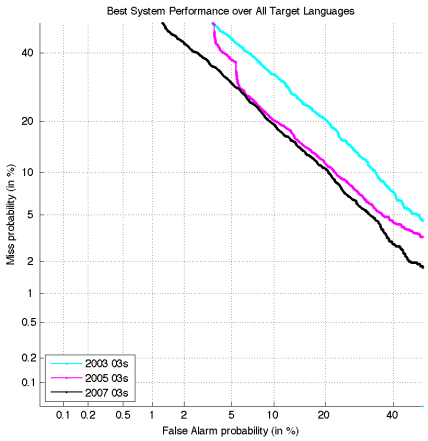

Figure 3: Best system performance on 30 second closed set general language recognition trials in the 2003, 2005, and 2007 evaluations

From Tables 3 and 4 and figures 1, 2, and 3 it is apparent that considerable progress in general language recognition performance was seen in 2007 compared with the results of earlier years. It should be noted that this occurred even as the 2007 evaluation included more target languages than the earlier years. But it should also be observed that the degree of performance improvement was greater for longer than for shorter duration trials.

Figure 4 presents DET plots comparing performance of the overall best systems for closed set general language recognition in 2005 and 2007 when the target language is restricted to a specific common target language of the evaluations. The curves represent language detection capabilities for English, Japanese, Korean, and Tamil in the two evaluations. The general trend of better performance in 2007, particularly for longer duration trials, may be observed. Note that 30-second duration performance curves for Japanese and Korean do not appear in the charts. For the best overall system and the limited number of target trials in these languages performance is literally "off the charts", with no DET points involving miss and false alarm rates in excess of 0.05%.

Figures 5 and 6 present DET plots of closed set English and Mandarin dialect recognition, respectively, for the best performing primary systems on these tests in the 2005 and 2007 evaluations. Figure 5 shows the DET curves specifically for American English as target dialect, while figure 6 shows the curve for Mainland Mandarin as target dialect. Since there are only two dialects included in these closed set tasks, the corresponding curves for the alternative dialects are symmetric to those presented.

Table 3 and figure 5 show little change in English dialect performance between 2005 and 2007. Performance seems to have improved a bit on 3 second duration segments and degraded a bit on 30 second segments. Further analysis may help explain this lack of performance progress.

Table 3 and figure 6 indicate some performance improvement from 2005 to 2007 in Mandarin dialect recognition, particularly on the longer durations. There was concern following the 2005 evaluation that the Mainland/Taiwan dialect distinction was not

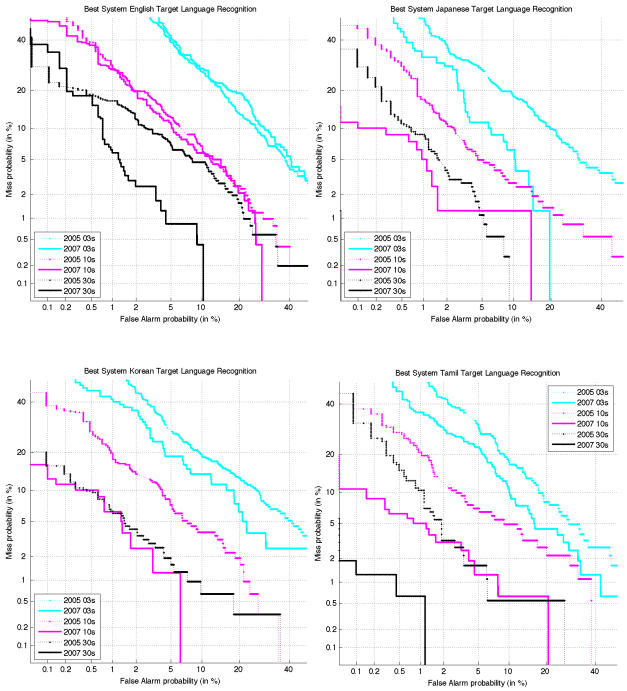

Figure 4: Overall closed set language recognition best system performance with target language restricted to (clockwise from the upper left) English, Japanese, Korean, and Tamil for each duration in 2005 (broken) and 2007 (solid). There is no 30-second duration curve for 2007 for Japanese and Korean because best system performance is so good as to be "off the chart"

precisely enough defined. Further analysis may attribute this year's improvement to better dialect auditing in preparing the test data.

The results in tables 3 and 4 for Hindustani and Spanish dialect recognition, which were not tested in recent past evaluations, appear rather disappointing. The validity of the dialect distinctions being investigated and the quality of the auditing for dialect may need to be discussed further. It may be enlightening

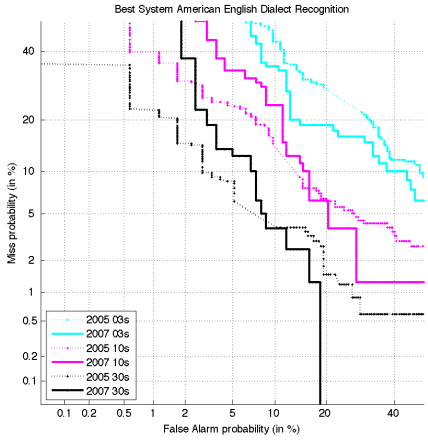

Figure 5: Best system performance for closed set American English dialect recognition for each duration in 2005 (broken) and 2007 (solid).

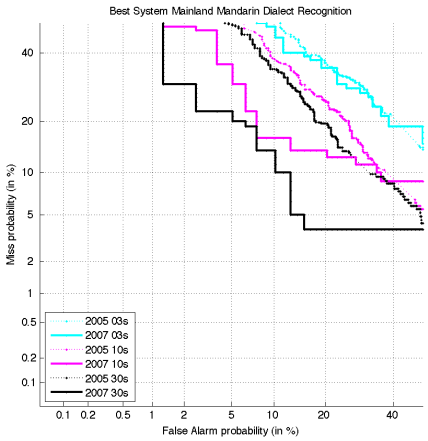

Figure 6: Best system performance for closed set Mainland Mandarin dialect recognition for each duration in 2005 (broken) and 2007 (solid).

to determine how well humans fluent in the languages involved can distinguish these dialect pairs. A limited experiment to do this is being undertaken.

7. Future plans

It is expected that future language recognition evaluations will be held at least once every two years. The next evaluation should occur in 2009.

8. References

[1] A. F. Martin and M. A. Przybocki, "1996 NIST Language Recognition Evaluation", NIST, Gaithersburg, MD [Online]. Availble: http://www.nist.gov/speech/tests/lang/1996

[2] A. F. Martin and M. A. Przybocki, "2003 NIST Language Recognition Evaluation", NIST, Gaithersburg, MD [Online]. Availble: http://www.nist.gov/speech/tests/lang/2003

[3] A. F. Martin and M. A. Przybocki, "NIST 2003 Language Recognition Evaluation", *Proc. EuroSpeech, 2003,* Geneva, Switzerland, Sep. 2003, pp. 1341-1344.

[4] A. F. Martin and A. N. Le, "2005 NIST Language Recognition Evaluation", NIST, Gaithersburg, MD [Online]. Availble: http://www.nist.gov/speech/tests/lang/2005

[5] A. F. Martin and A. N. Le, "The Current State of Language Recognition: NIST 2005 Evaluation Results", *Proc IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, PR, Jun. 2006.

[6] A. F. Martin and A. N. Le, "2007 NIST Language Recognition Evaluation", NIST, Gaithersburg, MD [Online]. Availble: http://www.nist.gov/speech/tests/lang/2007

[7] N. Brummer and J. du Preez, " Application-independent evaluation of speaker detection", *Computer, Speech & Language*, v. 20, issues 2-3, April-July 2006.

[8] N. Brummer and D. A. van Leeuwen, "On calibration of Language Recognition Scores", *Proc IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, PR, Jun. 2006.

[9] A. F. Martin et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. EuroSpeech 1997,* Rhodes, Greece, Sep. 1997, pp. 1985-1988.