Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

NIST Language Recognition Evaluation – Past and Future

Alvin F. Martin[1], Craig S. Greenberg[1], John M. Howard[1], George R. Doddington[1], John J. Godfrey[2]

[1]National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
{alvin.martin|craig.greenber|john.howard|george.doddington}@nist.gov, godfrey.jack@gmail.com

Abstract

This is a review of the six NIST Language Recognition Evaluations from 1996 to 2011. The evolving nature of the task is described, including the (non-)distinction between language and dialect. The languages/dialects tested are noted, and the challenges of data collection for such evaluations and the collections actually undertaken are reviewed. The performance measures employed are defined, and the performance levels achieved in both earlier and later evaluation tasks on different tests are discussed. Plans for the next evaluation in the series are presented.

1. Introduction

NIST has coordinated evaluations of automatic language recognition technology in 1996, 2003, 2005, 2007, 2009, and 2011 (see [1]). They are designated LRE96, LR03, etc. The next evaluation in this series is planned for late 2014 or early 2015.

The term language recognition may be somewhat vague both with respect to the meaning of "language" and to the meaning of "recognition". With regard to the former, the earlier evaluations sought to make a hard distinction between language recognition and dialect recognition, and had separate tests for each, with the latter involving distinguishing dialect pairs within the same language. This distinction was abandoned in the more recent evaluations, with the emphasis shifted to distinguishing pairs of language classes, whether they be called languages or dialects. (See the discussion of Section 3.)

The term recognition is a broad one. It may refer to identification, with the language of a speech segment to be determined from among a set of n specified languages. Alternatively, it may refer to detection, where for a language (or languages) of interest, the task is to determine whether or not the language is spoken in each speech segment, thus presenting a two-class problem. In either context, an "other" language class may be included, along with some test speech segments in unspecified additional languages, to make the task open-set in principle.

The NIST evaluations have emphasized the language detection viewpoint. In the earlier evaluations (through LRE09), for each target language, systems were asked to determine whether or not a given speech segment was of that target language. Given specified error costs and prior probabilities, the answer could be affirmative for multiple target languages for a given speech segment. The more recent shift to distinguishing language pairs largely eliminates the identification/detection issue.

Section 2 delves further into the overall history of the NIST LRE's, including the language classes included. Section 3 looks further at the distinction, or lack thereof, between languages and dialects. Section 4 discusses the key subject of evaluation data collection, and the problems and challenges it has presented. Section 5 looks at language recognition performance achieved in the earlier LRE's. Section 6 notes the changes adopted for the more recent evaluations and the performance measures adopted for them. Section 7 looks at LRE11 overall performance, and Section 8 looks at performance for particular language pairs of interest in recent evaluations. Section 9 describes the planned and soon to be conducted next LRE. We summarize in Section 10.

2. LRE History

Each LRE has included a list or lists of the target languages and dialects to be tested. For each target and each test speech segment, systems were asked to provide an actual decision ("true" or "false") and a score, with higher scores indicating greater likelihood that the target language class is present. Each evaluation also consisted of three groups of test segments based on duration, containing approximately 3, 10, or 30 seconds of speech.

LRE96 and LRE03 used twelve target languages, while LRE05 used a subset of these. For three of these twelve, there were two individual dialects included as separate tests in LRE96 or LRE05. Table 1 details this information.

Table 1: Languages/dialects included in LRE96, LRE03, and LRE05 (where 'X' indicates inclusion, and 'D' indicates use in a dialect test as well)

Language	'96	'03	'05	Dialects/Remarks
Arabic	X	X		Conversational Egyptian
English	D	X	D	Gen./Southern American in '96 American only in '03 American/Indian in '05
Farsi	X	X		
French	X	X		Canadian
German	X	X		
Hindi	X	X	X	
Japanese	X	X	X	
Korean	X	X	X	
Mandarin	D	X	D	Mainland/Taiwan in '96, '05
Spanish	D	X	X	Caribbean/Highland in '96 Latin American only in '03 Mexican only in '05
Tamil	X	X	X	
Vietnamese	X	X		

LRE07 took the two-tiered language/dialect tests of the prior evaluations one level further. In addition to dialect and general language recognition tests, four Chinese languages (sometimes called dialects) were included, namely Cantonese, Mandarin, Min, and Wu. Thus Chinese was a target language for general language recognition, the four named were targets for Chinese language recognition, and Mainland and Taiwan were targets for Mandarin

dialect recognition. Table 2 presents all the languages and dialects included.

Table 2: LRE07 target languages and dialects

Arabic	English	Farsi
Bengali	American	German
Chinese	Indian	Japanese
Cantonese	Hindustani	Korean
Mandarin	Hindi	Russian
Mainland	Urdu	Tamil
Taiwan	Spanish	Thai
Min	Caribbean	Vietnamese
Wu	non-Caribbean	

LRE09 encompassed the transition from the general language recognition task to the language pairs task. The former was the main (required) evaluation test, with the latter an optional task. There was no hierarchical separation of languages and dialects, but just a single list of "languages", some of which would have been described as dialects in earlier LRE's. However, eight pairs, some of which would have been viewed as dialect tests in the earlier evaluations, were designated as of particular interest. Table 3 lists the 23 languages and eight pairs.

Table 3: LRE09 target languages. The first two columns give the language pairs of particular interest

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

While the full language pairs test (276 such pairs) was optional and only completed by two participating sites, tests of some or all of the eight tests of particular interest were performed by a number of participants.

LRE11 was fully devoted to the language pairs task. All participants were required to provide decisions and scores for all 300 pairs involving the 24 specified target languages listed in Table 4. This table also shows the six clusters of related languages/dialects into which 19 of the targets could be grouped, along with the 5 other languages included. Much of the analysis of the results focused on performance within these six clusters.

3. The Meaning of Language, or Dialect

The popular adage has it that "a language is a dialect with an army and a navy".[1] This indeed often holds in general usage. (The languages of China are frequently referred to as dialects, while the term of distinction (if any) between Hindi and Urdu and between Serbian and Croatian has varied largely with whether their speakers were living in the same or different nation-states.)

Table 4: LRE11 target languages and clusters

Cluster	Classes
Arabic	Iraqi, Levantine, Maghrebi, Modern Standard
English	American, Indian
Indo-Aryan	Bengali, Hindi, Panjabi, Urdu
Persian	Dari, Farsi
Slavic	Czech, Polish, Russian, Slovak, Ukrainian
Tai	Lao, Thai
Other:	Mandarin, Pashto, Spanish, Tamil, Turkish

Mutual intelligibility is one criterion that is sometimes suggested as a basis for distinctions. But this is problematic, as intelligibility may be partial and need not be a symmetric, let alone a transitive, relationship. The notion of a dialect continuum is often noted.[2]

The term dialect is used in multiple, and sometimes conflicting, ways. It may refer to different forms of written or of spoken language, and for spoken language, may refer to differences in formal or in informal speech. (Chinese and Arabic language varieties are notably dependent of these distinctions.) It may refer to speech differences between different social classes, or different ethnic (or religious) groups, or different regional populations. At the regional level it may refer to broad differences between widely separated populations (e.g., British or American or Indian English) or to much finer geographic distinctions. (U.S. English is described as having as many as 24 different regional dialects.) At the opposite extreme from language differences is the notion of idiolect, corresponding to variety unique to a single person.

The NIST LRE evaluations have been more successful, in terms of performance results, and probably in terms of confidence in the ground truth auditing, with broader dialect class distinctions, most notably American/Indian English, than with narrower ones including Hindi/Urdu and Bosnian/Croatian. In view of the different types and levels of distinction that may be entailed in dialects, the choice in the more recent evaluations to simply have multiple language classes is probably to be preferred, and is likely to continue.

4. Issues of Data Collection

All speech processing technology evaluations are dependent upon the collection of appropriate data in sufficient quantity and variety, and language recognition poses some special problems and challenges. In particular, it is important that language variability not correlate with other factors characterizing the data being used. Thus it is problematic to collect data for the different target languages in different countries where they are each commonly spoken.

The earlier NIST LRE's were structured around collections of conversational telephone data. LRE96 and LRE03 used primarily speech segments selected from the phone conversations collected

[1] Max Weinrich attributes "*a shprakh iz a dialekt mit an armey un flot"* to an unidentified auditor at a lecture in 1943 or 1944 in his speech "The

[2] See http://en.wikipedia.org/wiki/Dialect_continuum. Some of the problems with the mutual intelligibility criterion are discussed in: http://ccat.sas/upenn.edu/~haroldfs/540/langdial/node2.html#SECTION00011000000000000000.

for the Linguistic Data Consortium's CallFriend Corpus [2]. This corpus consisted of conversation sides collected from people in the U.S. who agreed to be recorded in exchange for being able to make a free call to family members or friends. The published corpus provided development data, with test data selected from the conversations that were recorded as part of the collection process but withheld from the originally published corpus.

Corpora for language recognition need to have multiple speakers of each language; repeat speakers need to be avoided. But the capability to make free long distance calls home ceased to be a major benefit in the current millennium. Thus it became increasingly difficult and costly to collect corpora of phone calls where only a single call per speaker was desired.

LRE05 and LRE07 nevertheless were also implemented using solely conversational telephone data. LRE05 used primarily data collected by the Oregon Health and Science Univeristy, along with limited remaining CallFriend data. As noted in Table 1, the languages were a subset of those included in the prior evaluations.

LRE07, as noted in Table 2, expanded the set of language classes. The LDC, as part of its collection of the Mixer 3 Corpus [3] for speaker recognition (involving multiple languages), sought data to simultaneously support evaluation in both speaker and language recognition. One conversation collected from each solicited and compensated participant was to be used for language, while other conversations of the participant would be used for speaker recognition, where multiple conversations were needed.

The cost problem for telephone data led with LRE09 to a modified collection paradigm using narrowband broadcast data in addition to phone calls. Many radio and television broadcasts include a proportion of narrowband speech coming from telephone sources. These may include reporters in the field and people calling in to express opinions as is common in the talk radio genre.

For LRE09 a preliminary experiment was carried out using then state-of-the-art language recognition systems on human audited segments of narrowband speech from broadcasts in multiple languages by the Voice of America [5]. The results suggested that performance on these segments was comparable to that on conversation telephone segments with similar total speech durations within them. With a large quantity of voice VOA data, including some audited narrowband segments supplied for system development, LRE09 then used primarily segments from Voice of America broadcasts as test data (limited amounts of available phone conversation data was also included for comparison).

For LRE11, the LDC initiated a major new collection effort specifically for LRE purposes, using a hybrid of the previous collection approaches. The LDC sought new data from both phone conversations and broadcast sources.

Phone data was collected using "claques". A claque involved a recruited native speaker of a language class in the U.S. initiating a phone conversation with each of a number of other native speakers in his/her circle of acquaintances. The claque leaders thus had multiple conversations but they were not used in the test data; rather the other conversation sides of their calls were utilized. Multiple claques were recruited for each language class.

Meanwhile, multiple broadcast sources with narrowband speech were sought for each target class. A diversity of sources with only a few segments from each individual show was perferred. This strategy was designed to limit the numbers of repeat speakers for each target class.

For most languages a combination of phone and broadcast segments was produced, but the mix varied among languages. In the case of the Arabic varieties, Modern Standard Arabic came only from broadcast sources, while for the regional Arabic dialects all data came from phone conversations.

Another change in LRE11 was in the data format. Prior evaluations provided 8-bit ulaw data, as appropriate for (U.S. based) telephone calls, but artificial for narrowband broadcast collections. For LRE11 it was decided that all data, which came originally from multiple sampling rates and sample sizes, would be converted to 16-bit linear pcm.

Table 5 summarizes the corpora used in the six LRE's to date and the types of speech collected for each. Further, it shows the (approximate) total number of segments collected for each duration and the format of the data provided. Note that the switch to primarily broadcast data for LRE09 allowed a big increase in the number of test segments, as well as in the number of target classes.

Table 5: Data Sources for the NIST LRE's

LRE	Corpus Source	Speech Type	Total Segments per Duration	Format
96	CallFriend	CTS	~1500	8-bit ulaw
03	CallFriend	CTS	1280	8-bit ulaw
05	OGI	CTS	3662	8-bit ulaw
07	Mixer 3	CTS	~2500	8-bit ulaw
09	VOA	BNBS	~12000	8-bit ulaw
11	New LRE11 Corpus	CTS/ BNBS	~10000	16-bit linear pcm

5. Performance in Earlier Evaluations

The basic closed-set language detection task (see section 1) for each of three durations was included as primary task from 1996 to 2009. Detection performance was examined for each target language, and the primary performance metric was defined as

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

with the cost parameters C_{Miss} and $C_{FalseAlarm}$ set to 1 and the prior probability P_{Target} of a target trial always set to 0.5.[3]

Performance in these evaluations was generally viewed as quite good, with the best systems' levels of performance generally improving over successive evaluations. Figure 1 [4] summarizes best system performance on the basic closed set recognition task over all target languages over the course of the five evaluations. Performance generally improved steadily, with some plateauing for the longer durations.

Figures 2-5 [6] present DET (Detection Error Tradeoff) plots [7] showing recognition performance improvement for several specific target languages included in both LRE05 and LRE07. Note the lack of 30 s curves for LRE07 for Japanese and Korean, as the perfect performance falls off the lower left corner of the chart.

[3] For LRE96 and LRE03 the overall metric was specified as in equation (1) as the mean of the miss and false alarm rates across all target languages, while for subsequent LRE's this metric was computed for each target, and these were then averaged, but this makes little overall difference.

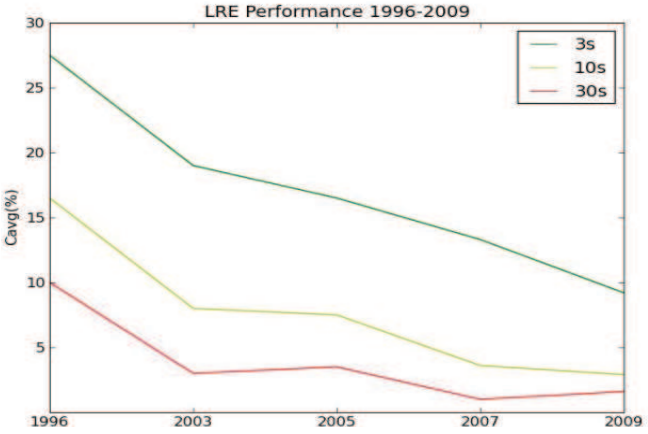

Figure 1: Best system closed-set overall scores for the NIST LRE evaluations 1996-2009

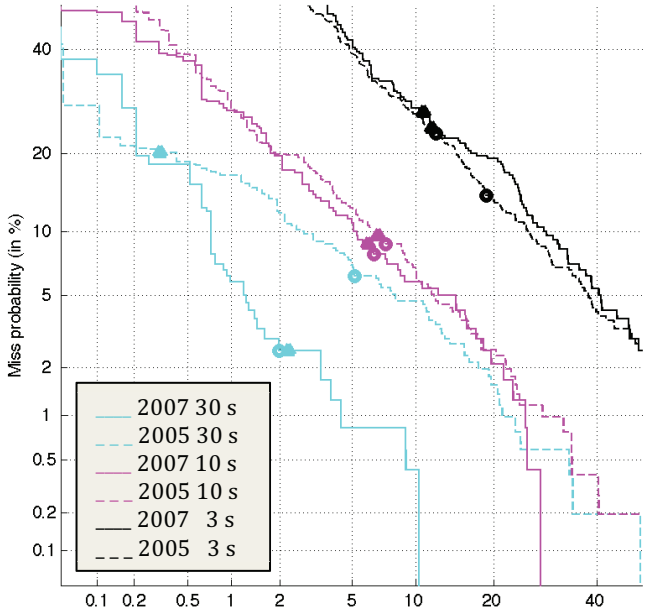

Figure 2: Closed set language recognition best system performance with target language restricted to English, for the three durations in LRE05 (broken) and LRE07 (solid).

6. Re-specifying the Task

The sense of steady improvement from LRE96 to LRE09 on what was already viewed as quite good performance on the basic language detection task, with results approaching perfection for longer duration segments of some distinct language classes, contributed to the decision to shift the evaluation focus. Given evidence of a strong capability to distinguish major language classes from dissimilar others, the view developed that the primary challenge and interest lay in distinguishing specific language pairs, particularly ones involving closely related language classes.

The general language pairs task was offered and encouraged in LRE09, and became the sole task in LRE11, as it will be in a modified form in the upcoming LRE. In these two most recent evaluations, systems were asked for each of the thousands of test segments and each of the hundreds of language pairs, to specify with a decision and score which language class of each pair corresponded to each segment, given that the segment contained one of the language classes of the pair. Scoring was done only for pairs that included the language of each segment; all other submitted results were ignored.

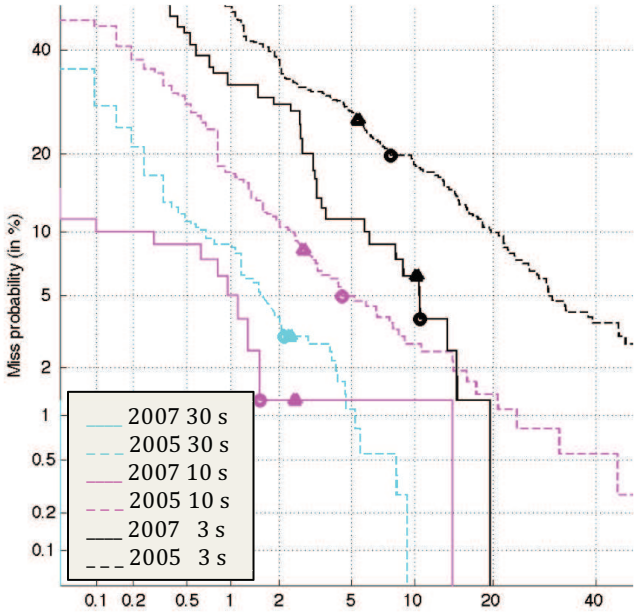

Figure 3: Closed set language recognition best system performance with target language restricted to Japanese, for the three durations in LRE05 (broken) and LRE07 (solid). For LRE07 30 s, performance is "off the chart".

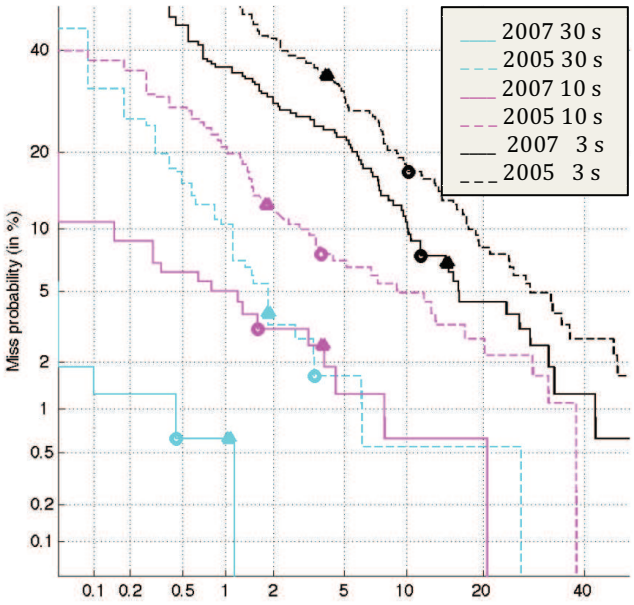

Figure 4: Closed set language recognition best system performance with target language restricted to Tamil, for the three durations in LRE05 (broken) and LRE07 (solid).

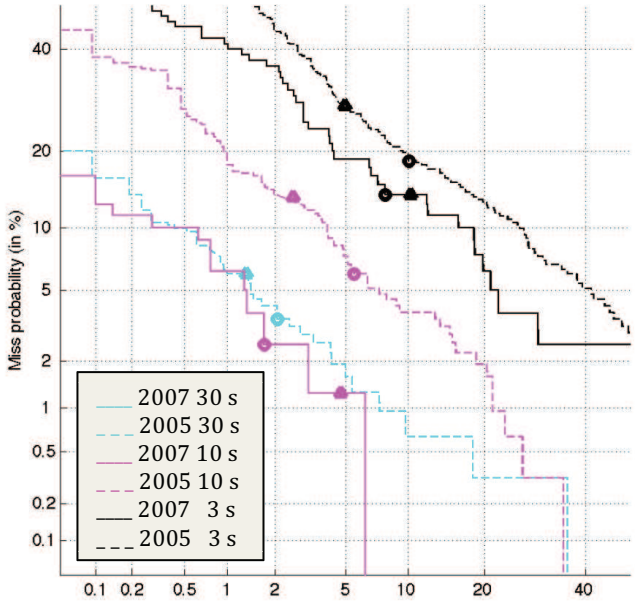

Figure 5: Closed set language recognition best system performance with target language restricted to Korean, for the three durations in LRE05 (broken) and LRE07 (solid). For LRE07 30 s, performance is "off the chart".

For the language pair task, scoring was based on a simple language pair cost function defined for each language pair (L_1 / L_2) as

$$C(L_1, L_2) = C_{L_1} \cdot P_{L_1} \cdot P_{Miss}(L_1) + C_{L_2} \cdot (1 - P_{L_1}) \cdot P_{Miss}(L_2) \quad (2)$$

Here C_{L_1}, C_{L_2} and P_{L_1} are viewed as application parameters representing the costs of the two error types and the prior probability of L_1, respectively. They were assigned the symmetric values $C_{L_1} = C_{L_2} = 1$, and $P_{L_1} = 0.5$, thus making the cost function the mean error rate.

This measure was then computed separately for each of the three segment durations, and for both actual and (score threshold based) minimum cost decisions. The difference between the actual and minimum cost may be viewed as the system's calibration error.

As noted, two LRE09 participants performed testing on all 276 language pairs. Examination of one set of these results for 30 s duration segments shows that the two most confusable pairs, by far (with mean error rate in excess of 25%), were Hindi/Urdu and Bosnian/Croatian. These are notably pairs whose distinctness is based more on geopolitical than on linguistic factors. Russian/Ukrainian was next (mean error rate around 11%), and only eleven pairs had a mean error rate in excess of 1%. These included six of the eight pairs of particular interest. The exceptions were the largely not mutually intelligible pairs Portuguese/Spanish[4] and Cantonese/Mandarin, (For the latter, and perhaps the former as well, the spoken languages differ considerably more than the written forms.)

[4] On Spanish and Portuguese see, for example, John B. Jensen, "On the Mutual Intelligibility of Spanish and Portuguese", Hispania, Vol. 72, No. 4, Dec. 1989

These results point to the outsize role among total errors played by the particularly difficult language pairs involving fairly closely related language classes as well as the excellent level of overall performance for most of the possible language pairs. For LRE11, this required reconsideration of how the main overall evaluation metric should be defined. A simple average across all language pairs did not seem appropriate.

Instead, an average over only the most difficult pairs seemed a better choice. Rather that choosing a list of such pairs for the 24 language classes of LRE11, we chose to have the cost function for each system be the mean cost over the 24 pairs that proved most difficult for that system. Specifically, the 24 pairs for which the minimum cost operating points for 30 s duration segments were greatest were determined for the system. For each duration, the system's official performance measure was the mean of the 24 actual decision cost function values for these pairs.

This would mean that calibration errors should not affect the choice of the 24 pairs to average, but would contribute to the measured performance based on these pairs.

7. LRE11 Overall Performance

Figure 6 (from [8]) shows the official overall performance results of the primary systems of LRE11, ordered by results on 30 s. These results were in line with expectations, and LRE11 was viewed as representing a successful implementation of the new evaluation paradigm.

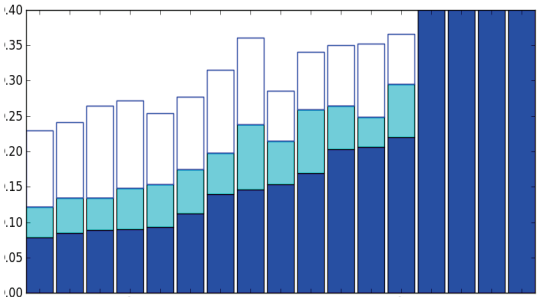

Figure 6: Overall performance measures for the 17 primary systems in LRE11. Each bar represents a system's performance, with the blue showing performance cost on 30 second segments, the blue + cyan on 10 seconds, and the blue + cyan + white on 3 seconds.

It may be noted, however, that Figure 6 presents comparisons of system scores over different sets of language pairs. We also examined [8] the outcomes if a common set of language pairs, based on using the hardest pairs for six leading systems, and the results were not dramatically different. There was, not surprisingly, a great deal of commonality in which were the most difficult pairs.

We expected the most difficult pairs to be from within the clusters listed in Table 4, and for the most part this was the case. The exceptions were a number of instances of non-cluster pairs involving the languages Pashto or Bengali (or both) among those that were found most confusable by the leading systems. It appears that there may have been some issues with the quality of the data collected for these two languages.[5]

[5] George Doddington did some listening and found that quite a few of the Pashto segments contained chanting rather than ordinary cadence speech.

The use of a metric based on which pairs proved most difficult for a system under certain conditions also opened up some possibilities for game playing. One site – it has been policy not to publicly associate participant names with their performance results – submitted an alternate system in which the 30 s duration minimum operating point results appeared to have been manipulated to produce unexpectedly poor results for certain dissimilar language pairs. Organizations coordinating evaluations need to be alert to such anomalous possibilities, and to establish rules and procedures to avert them.

8. Recent Performance on Particular Pairs

As noted, the language pairs task has received increasing emphasis in the past two or three LRE's. Here we examine best system performance for several specific language pairs which were included in both LRE09 and LRE11, and possibly in LRE07 as well. The charts show the best minimum score as defined above for the pair of interest for each of the three durations.

Figure 7 presents best system minimum scores for the American/Indian English pair for the past three LRE's. English dialect recognition has a history going back to LRE96, and likely has received the most attention by the (largely English speaking) participants in the evaluations. With the caveats noted about varying data sources across evaluations, these results suggest good performance improvement for this pair over the three evaluations.

Figure 8 presents best system minimum scores for the Hindi/Urdu pair for the past three LRE's. Here there is less strong evidence of progress over the course of these evaluations. As has been noted, the language/dialect distinction here is a problematic one and overall performance levels, especially for the shorter durations, is not at all impressive. A human test in one evaluation cycle also showed some issues about consistency with annotator judgment, so the value of pursuing this test pair is questionable.

Figure 9 presents best system minimum scores for the Dari/Farsi pair for the past two LRE's. Some improvement is seen for 30 and 3 second durations, with little change for 10 seconds.

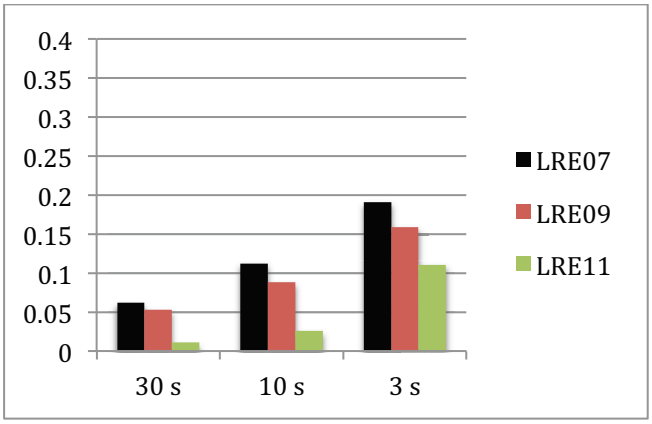

Figure 7: Best system minimum scores for American/Indian English in LRE07, LRE09, and LRE11 for 30, 10, and 3 s durations.

Figure 10 presents best system minimum scores for the Russian/Ukrainian pair for the past two LRE's. Here the best system performance distinctively declined from 2009 to 2011. This is disappointing, and perhaps it is simplest to suppose here a true difference in the data sources used.

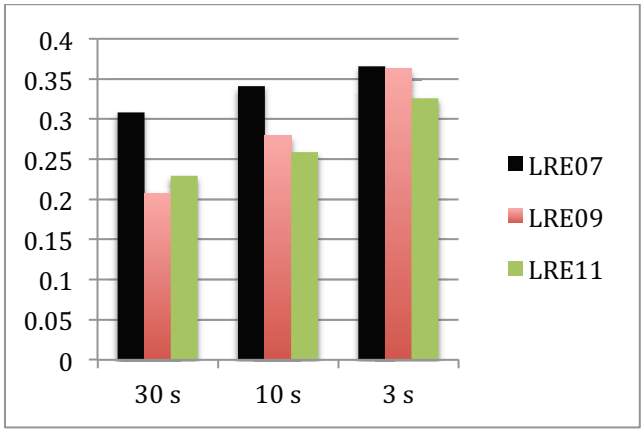

Figure 8: Best system minimum scores for Hindi/Urdu in LRE07, LRE09, and LRE11 for 30, 10, and 3 s durations.

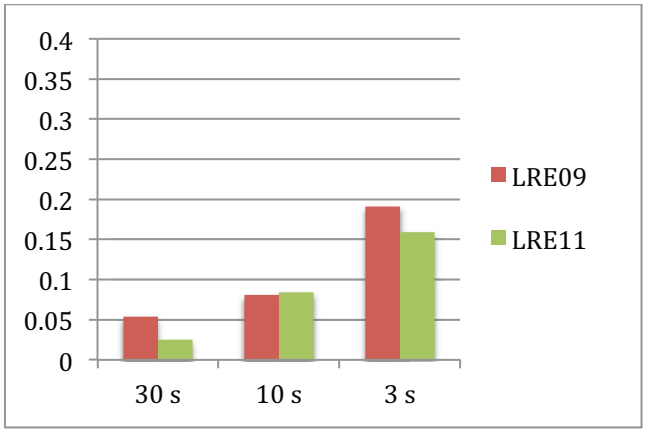

Figure 9: Best system minimum scores for Dari/Farsi in LRE09 and LRE11 for 30, 10, and 3 s durations.

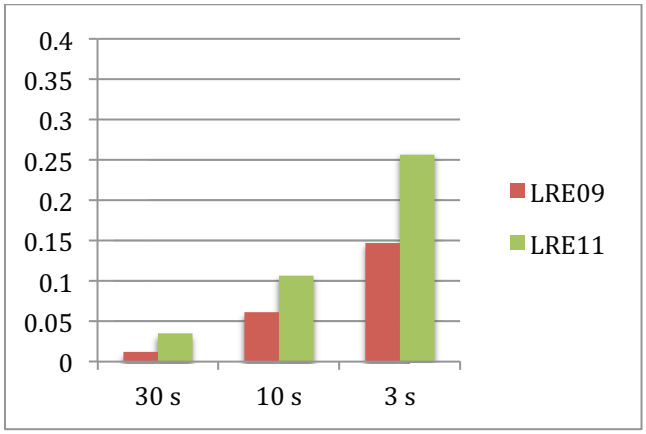

Figure 10: Best system minimum scores for Russian/Ukrainian in LRE09 and LRE11 for 30, 10, and 3 s durations.

9. Planning the Next Evaluation

The upcoming 2015 evaluation will concentrate on the task of distinguishing closely related language class pairs. The LDC is collecting speech data from six different language clusters, with two to five language/dialects included in each cluster. The resulting 20 classes are summarized in Table 6.

The data is being collected in a manner similar to that of the two preceding evaluations. There will be a mix of telephone call and narrowband broadcast speech. For Arabic, the MSA will be entirely broadcast, while the other varieties will be entirely from phone calls; for most other classes there will be a mix of telephone calls and narrowband broadcast data.

Table 6: Language clusters for the next LRE

Cluster	Classses
Arabic	Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
Chinese	Cantonese, Mandarin, Min, Wu
English	British, General American, Indian
French	West African, Haitian Creole
Slavic	Polish, Russian
Spanish	Caribbean, European, Latin American, Brazilian Portuguese

For each of these six clusters, there will be opportunities to compare pair performance with those obtained previously in LRE07, LRE09, or LRE11.

The scoring for each pair will use the same actual decision and minimum cost function as in recent evaluations. However, to simplify the task for participants, instead of requiring scores and decisions for each pair, systems will submit a 20-entry log likelihood vector for each test segment. Thus for such a submitted test segment vector L, for classes i and j in the actual cluster of the segment, the score for the pair i/j will be taken as $L_i - L_j$, with 0 as the actual decision threshold. Scoring will not be performed across clusters.

For each cluster, an overall cluster performance cost will be computed as the mean of performance costs over all n*(n-1)/2 class pairs within the cluster, where n is the number of classes in the cluster. An overall performance costs will be computed as the mean of those for the six clusters. The actual decision overall mean will be the official overall evaluation metric.

10. Summary

NIST has coordinated six language recognition evaluations since 1996. They have all concentrated on the task of detecting target language classes of interest.

The nature of the language classes of interest has varied over time, however. Earlier evaluations achieved very high performance for classes distinct from one another, and had separate tests for less distinct classes described as dialects. The evaluations have moved away from the dialect/language distinction and toward a concentration on distinguishing closely related language classes in a pair-wise context.

The next evaluation is planned for late 2014. It will include twenty language classes with pairwise evaluation within six clusters of related languages. It will utilize both conversational telephone and broadcast narrowband speech collected by the LDC in a new effort similar to that used for LRE11. Participation is open to all who are interested in the challenge.

11. Disclaimer

These results are not to be construed or represented as endorsements of any participant's system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials may be identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

12. References

[1] NIST, "Language Recognition Evaluation". Online: http://www.nist.gov/itl/iad/mig/sre.cfm

[2] Mark Liberman, Christopher Cieri *The Creation, Distribution and Use of Linguistic Data* LREC 1998: 1st International Conference on Language Resources and Evaluation, Granada, Spain, May 28-30 1998

[3] Christopher Cieri, Linda Corson, David Graff, Kevin Walker *Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora* Interspeech 2007: 10th International Conference on Spoken Language Processing, Antwerp, August 27-31 2007

[4] Christopher Cieri, et al., "The Broadcast Narrow Band Speech Corpus: A New Resource Type for Large Scale Language Recognition", *Proc. Interspeech 2009*, Brighton, UK, September 2009

[5] Alvin Martin. and Craig Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Proc. Odyssey 2010*, Brno, Czech Republic, 2010.

[6] Alvin Martin, and Audrey Le, "NIST 2007 Language Recognition Evaluation", *Proc. Odyssey 2008: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 2008

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

[8] Craig Greenberg, Alvin Martin, and Mark Przybocki, "The 2011 NIST Language Recognition Evaluation", *Proc. Interspeech 2012*, Portland, Oregon, USA, Sep. 2012