Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

The 2009 NIST Language Recognition Evaluation

Alvin Martin, Craig Greenberg

National Institute of Standards and Technology
Gaithersburg, Maryland, USA

alvin.martin@nist.gov, craig.greenberg@nist.gov

Abstract

This paper reviews the 2009 NIST Language Recognition Evaluation (LRE09), the most recent in a series held since 1996, which have evaluated automatic systems for language recognition. The 2009 evaluation was notable for including a larger number of target and non-target languages, for primarily utilizing "found" narrowband conversational broadcast data from the Voice of America, and for including a language pairs test condition that included examination of performance at distinguishing several particularly interesting and confusable pairs of languages. Overall, the broadcast data proved roughly comparable in difficulty with the type of collected conversational telephone date utilized previously. Improvement was seen in best system performance levels for some test conditions.

1. Overview

Successful evaluations of speech processing technology depend crucially on obtaining appropriate quantities of evaluation test data. Doing so in a cost effective manner is particularly challenging for language recognition evaluation, where the need is to collect conversational speech samples from large numbers of speakers in each of numerous languages without having language related differences in the collection channels.

Prior NIST Language Recognition Evaluations relied on paying subjects within the United States to engage in a single telephone conversation with a fellow speaker in their common native language. The decreasing cost of telephone access has made it increasingly difficult and expensive to induce people to engage in a project involving a single phone call. At the same time there has been increasing desire to include in the evaluation more languages and dialects, and to include a larger number of samples and speakers of each target language.

For the 2009 NIST evaluation the decision was made to go to a different data collection protocol involving "found" data. Narrowband data from previously collected sets of Voice of America (VOA) broadcasts were selected to provide large quantities of comparable data in a broad range of languages.

This collection was assembled by the Linguistic Data Consortium (LDC), and its feasibility for use in language recognition was investigated by researchers at the Brno University of Technology. See [1]. It provided the greater part of the data used in the 2009 NIST Language Recognition Evaluation (LRE09). Some conversational telephone data segments that had been collected for previous evaluations, but not used in them, were also included in LRE09, allowing a comparison of system performance on the different data types (see section 9 below).

The sections below review the protocols of LRE09 and the performance results obtained. Best LRE09 results are compared with those of previous NIST language recognition evaluations. Many of the performance charts presented here, and others as well, are available on the NIST web site [2]. Further information on the earlier evaluations is available on the NIST web site [3] and also in [4], [5], and [6].

2. Test Conditions

The basic evaluation task in LRE09 was the same as in prior evaluation: Given a segment of speech and a hypothesized target language, determine whether or not the target language is spoken in a given test speech segment, based on automated analysis of the data contained in the segment. A hard decision (yes or no) and a score (higher scores indicating greater confidence of a yes decision) were required for each such trial.

There were 23 target languages included in LRE09, as specified in Table 1. Note that in 2009 there were no separate dialect tests, and that certain pairs that in prior evaluation were distinguished only in separate dialect tests, including American English-Indian English and Hindi-Urdu, were separately included in the set of target languages in 2009.

Table 1: The 23 target languages.

Target Languages	
Amharic	Hindi
Bosnian	Korean
Cantonese	Mandarin
Creole (Haitian)	Pashto
Croatian	Portuguese
Dari	Russian
English (American)	Spanish
English (Indian)	Turkish
Farsi	Ukrainian
French	Urdu
Georgian	Vietnamese
Hausa	

Three different alternative hypothesis test conditions were included in LRE09:

1. For each trial, the set of non-target languages consisted of the languages in Table 1, minus the target language. This "closed-set" test condition was required of all LRE09 participants.

2. For each trial, the set of non-target languages included those of 1 above, plus other "unknown" languages whose identities were not disclosed. This "open-set" test condition was optional.
3. For each trial, the set of non-target languages consisted of a single language. This "language-pair" test condition was optional.

The closed and open-set conditions were similar to conditions included in prior evaluations. The language-pair condition was new, and is believed to be appropriate for some applications of interest. Participants could choose to do any or all of the 253 possible pairs. Eight specific pairs, however, were designated pairs of particular interest that participants were encouraged to attempt. These eight, which include pairs treated as dialect tests in prior evaluations, are listed in Table 2.

Table 2: The eight language pairs of particular interest.

Language Pairs of Interest
Bosnian-Croatian
Cantonese-Mandarin
Creole (Haitian)-French
Dari-Farsi
English (American)-English (Indian)
Hindi-Urdu
Portuguese-Spanish
Russian-Ukrainian

3. Data

Evaluation participants were offered all of the conversational telephone speech (CTS) data from the four past LRE's to develop and train their systems and, for languages not included in prior LRE's, a limited quantity Voice of America data that had been human annotated (by the LDC) along with a large quantity of Voice of America data labeled by language via an automatic process (possibly errorful). The number of such human annotated 30-second segments was close to 200 for most of the 12 languages involved, and as low as 142 for a few languages with limited data availability. Participants could also utilize additional data from publicly available sources. System descriptions were also required to document any outside data used.

The evaluation data consisted of human annotated segments drawn from Voice of America broadcasts along with, for some languages, segments of CTS data left over from that collected for prior evaluations. There were segments of approximately 30, approximately 10, and approximately 3 seconds of speech, with the 10-second segments contained within 30-second segments, and the 3-second segments contained within 10-second segments.

Table 3 lists for each of the 23 target languages and for the 16 out-of-set test languages included in the evaluation the numbers of human annotated 30-second training segments provided and the numbers of 30-second test segments included from VOA and from leftover CTS data

Table 3: The number of 30-second VOA train segments and VOA/CTS test segments by language.

Language	VOA Train	VOA Test	CTS Test
Amharic	171	398	- - - - -
Bosnian	194	355	- - - - -
Cantonese	- - - - -	62	316
Creole-Haitian	186	323	- - - - -
Croatian	181	376	- - - - -
Dari	194	389	- - - - -
English-Am.	- - - - -	374	522
English-Ind.	- - - - -	- - - - -	574
Farsi	- - - - -	338	52
French	196	395	- - - - -
Georgian	142	399	- - - - -
Hausa	200	389	- - - - -
Hindi	- - - - -	397	270
Korean	- - - - -	318	145
Mandarin	- - - - -	390	625
Pashto	197	395	- - - - -
Portuguese	166	397	- - - - -
Russian	- - - - -	254	257
Spanish	- - - - -	385	- - - - -
Turkish	194	394	- - - - -
Ukrainian	194	388	- - - - -
Urdu	- - - - -	347	32
Vietnamese	- - - - -	27	288
Arabic	Out-of-set	187	- - - - -
Azerbaijani	Out-of-set	366	- - - - -
Belorussian	Out-of-set	363	- - - - -
Bengali	Out-of-set	- - - - -	43
Bulgarian	Out-of-set	375	- - - - -
Italian	Out-of-set	- - - - -	30
Japanese	Out-of-set	- - - - -	180
Punjabi	Out-of-set	- - - - -	9
Romanian	Out-of-set	400	- - - - -
Shanghai-Wu	Out-of-set	- - - - -	69
Southern-min	Out-of-set	- - - - -	48
Swahili	Out-of-set	396	- - - - -
Tagalog	Out-of-set	- - - - -	84
Thai	Out-of-set	- - - - -	188
Tibetan	Out-of-set	368	- - - - -
Uzbek	Out-of-set	382	- - - - -

4. Performance Measurement and Representation

The performance of a detection system is characterized by its miss and false alarm error rates. LRE09 utilized a decision cost function that equally weights the two error types. An overall *average decision cost function* C_{avg} was then defined for each test condition by averaging over all languages included, as detailed in Figure 1.

NIST has traditionally used Detection Error Tradeoff (DET) curves [7] to represent the range of possible system operating points of detection systems, and plots of such curves

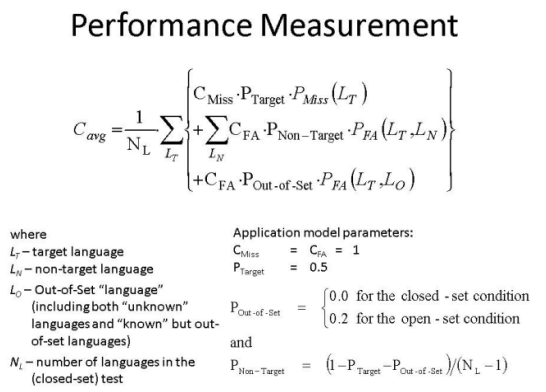


Figure 1: Decision cost function used as performance measuremt metric

are included below. But some investigators question the propriety of presenting DETs that pool multiple trials of test segments with different target languages. Since the total set of possible language classes for a segment is a (relatively small) finite number, there is a lack of independence across trials.

The strange effects this may cause is most apparent in the language pairs situation where there are only two possible target languages. Figure 2 offers an example involving Russian and Ukrainian. Note that the two single target language DET curves are necessarily symmetric. But the pooled curve can end up showing much worse performance that either of these, as occurs for System 2 in the figure. The differing shapes of the score histograms for the two languages, also shown in the figure, accounts for this.

5. Participants

A diverse group of organizations from four continents participated in LRE09. Table 4 lists in alphabetical order most of these.

Note that it has been NIST policy not to publicly identify participating sites with their evaluation performance results. Sites are permitted to discuss their own results but not, without permission, those of other participating sites.

6. Overall Results

The figures in this section review the overall performance results for the primary systems in LRE09.

Figure 3 presents the C_{avg} scores of primary systems for the three test conditions and the three test segment durations. Scores for each condition are presented via cumulative stacked bar charts over the durations. Thus, for example, the 3-second scores are represented by the total height of the three stacked rectangles. As expected, scores are always larger (worse) for shorter durations, though in some cases the 10-second scores are not far different from the 30-second ones.

Figure 4 displays the DET plots for the primary systems on 3-second test segments for the closed-set and open-set test conditions. They are generally quite linear, suggesting underlying normal distributions. The corresponding plots for 10 and 30 seconds may be viewed online [2].

Table 4: Organizations participating in LRE09.

Organization	Location
Universidad Autonoma de Madrid	Madrid, Spain
Brno University of Technology	Brno, Czech Republic
Agnitio	Somerset West, South Africa
Institute of Automation, Chinese Academy of Sciences	Beijing, China
Chinese University of Hong Kong	N.T., Hong Kong
University of the Basque Country	Bizkaia, Spain
iFlyTek Speech Lab, EEIS University of Science and Technology of China	HeFei, AnHui, China
Institute for Infocomm Research	Singapore
Institute of Acoustics, Chinese Academy of Sciences	Beijing, China
L2F-Spoken Language Systems Lab INESC-ID Lisboa	Lisbon, Portugal
Laboratorie Informatique D'Avignon	Avignon, France
CNRS-LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur)	Orsay, France
Loquendo	Torino, Italy
Politecnico di Torino	Torino, Italy
MIT Lincoln Laboratory	Lexington, MA, USA
National Taipei University of Technology, Department of Electrical Engineering & Graduate Institute of Computer and Communication Engineering	Taipei, Taiwan
Tsinghua University Department of Electrical Engineering	Beijing, China
Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek	Soestenberg, The Netherlands

Only two sites did all of the language-pair tests. Figure 5 (created by George Doddington) displays the C_{avg} scores for the pairs with the highest such scores on 30-second segments for one of these sites. The most confusable pairs were Hindi-Urdu and Bosnian-Croatian, which both are arguably language distinctions based as much on political as on linguistic factors. (See, for example [8] and [9]. These were followed by Russian-Ukrainian, American English-Indian English, and Dari-Farsi, all expected to be difficult pairs and included on the list of eight pairs of particular interest. Haitian Creole-French is also among the 18 with an overall error rate above

1%. It should be noted that two of the pairs designated as of particular interest, Cantonese-Mandarin and Portuguese-Spanish are not among these and were distinguished at a better than 99% rate for 30-second segments.

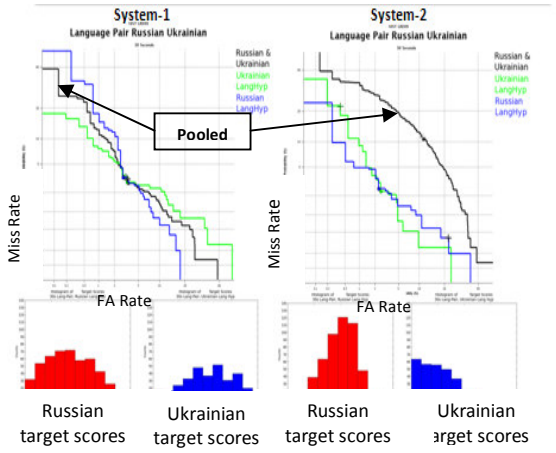


Figure 2: DET plots for two systems for the Russian-Ukrainian language pair. The pooled DET lies between the two single target language DETs for System 1 but not for System 2. The differing shapes of the System 2 target trial score histograms by language account for this phenomenon.

7. Comparison With Prior Evaluations

Figure 6 presents information on the history of the NIST LRE's with respect to the numbers of participants and the numbers of languages, target and out-of set included. A key objective of the 2007 and 2009 evaluations was to increase the numbers of languages represented, including out-of-set languages, as this is representative of some real-world application scenarios. These two most recent evaluations have also seen an increased level of participation compared with earlier years, though participation in 2009 was down from 2007, perhaps due to the challenges of processing the large VOA datasets provided.

Figure 7 examines performance history over the course of the NIST LRE's as represented by the best closed-set system C_{avg} scores in each evaluation for 30, 10, and 3-second segments. For 2009 the general downward trend continued for 3-second segments, while performance appears to be leveling off for the much better performing longer segments, with the 30-second best score actually increasing slightly.

Figures 8 and 9 presents DET plots comparing best system performance in 2007 and 2009 on segments of each of the three durations for the closed and open-set test conditions. (For 30 seconds, two best systems are included for LRE09.) For the open-set condition little difference is seen for the 30 and 10 second segments, with somewhat improved performance seen in 2009 for 3 second segments.

Figure 10 compares best system performance in 2007 and 2009 for closed-set recognition of six specific target languages common to the two evaluations, with respect to each duration. For most, better performance is seen in 2009. (The 2009 30-second Korean performance is close enough to perfect to be off-the-chart.) The exception is Russian, which followed the overall pattern of improved 2009 performance on 3-second segments but degraded performance on 30-second segments.

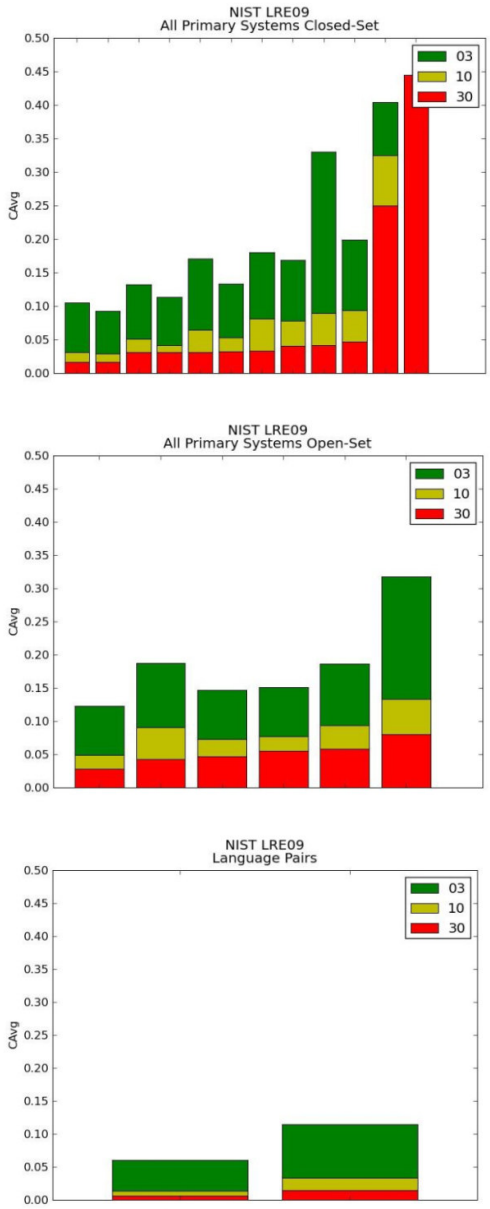
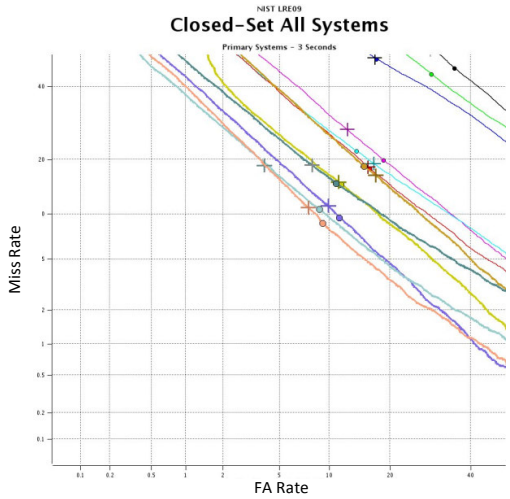


Figure 3: Bar charts display overall C_{avg} scores of primary systems on the closed-set, open-set, and language-pairs test conditions for 30, 10, and 3-second segments. Note that scores represented are cumulative over the durations.

Figure 11 compares best system language pairs results (previously called dialect tests) in 2007 and 2009 for the American English-Indian English and Hindi-Urdu language pairs. The plots are specifically for detecting American English in the context of the former pair and Hindi in the context of the latter pair. (This avoids the pooling anomaly illustrated in figure 2). For the former, improved performance is observed in 2009 for detecting American English for all durations. For the latter, 2009 improvement is seen for 30 and 10-second durations, while the 3-second challenge remains, with performance little better than random.



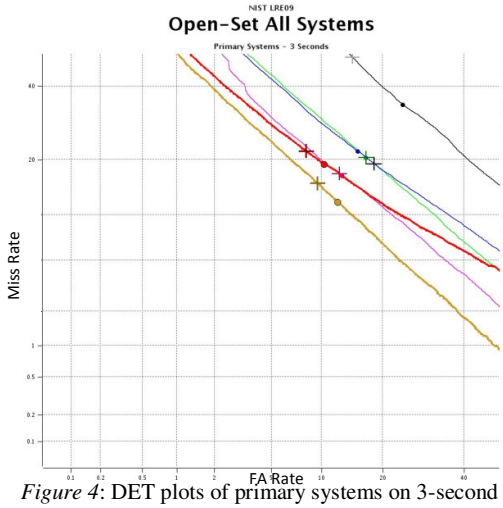


Figure 4: DET plots of primary systems on 3-second test segments for the closed-set and open-set test conditions.

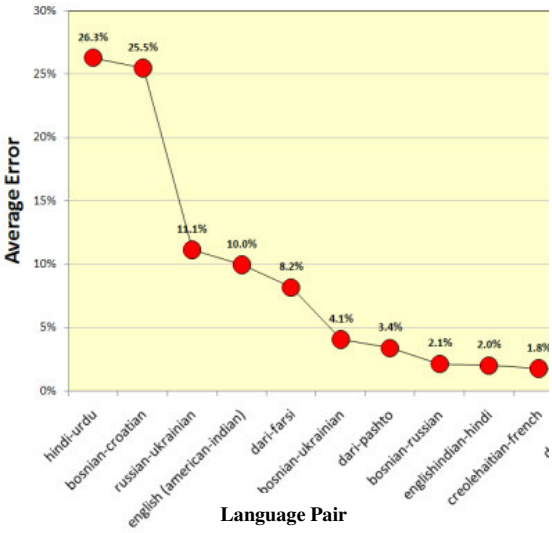


Figure 5: Scores for most confusable language pairs for one system on 30-second segments.

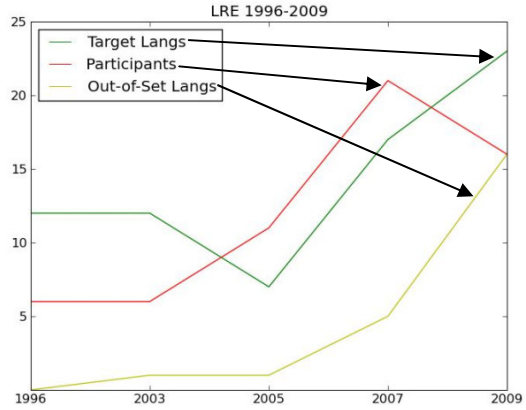


Figure 6: Numbers of languages and participating sites in the NIST LRE evaluation 1996-2009.

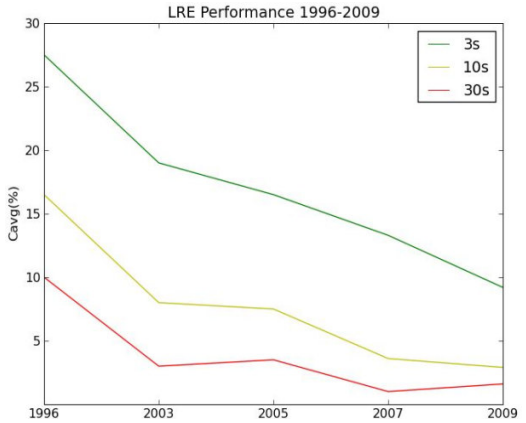


Figure 7: Best system closed-set C_{avg} scores in the NIST LRE evaluations 1996-2009.

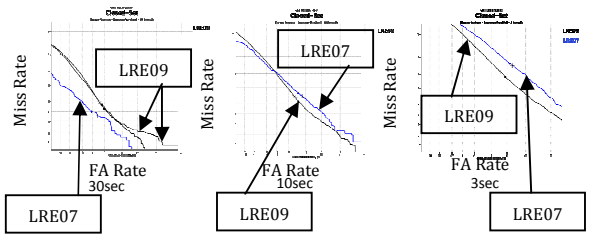


Figure 8: Best system DET plots for LRE07 and LRE09 for the closed-set test condition.

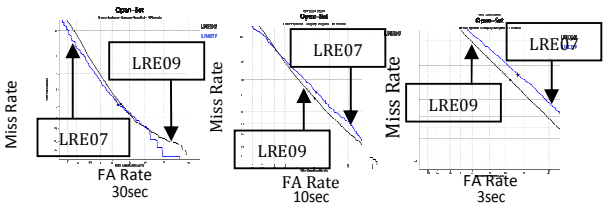


Figure 9: Best system DET plots for LRE07 and LRE09 for the open-set test condition.

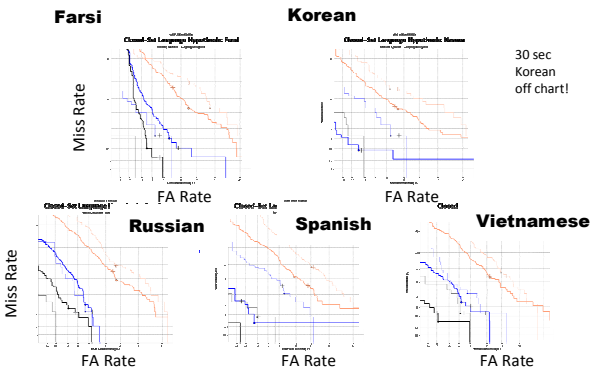


Figure 10: Best system DET plots for LRE07 and LRE09 for recognition of six target languages common to the evaluations for each duration.

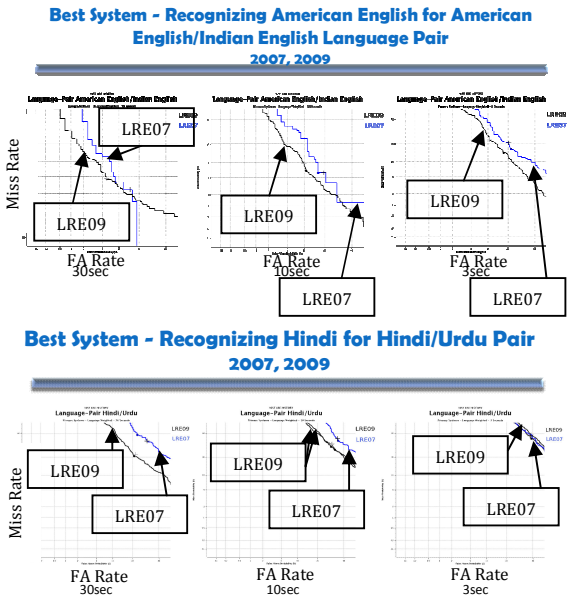


Figure11: Best system DET plots for LRE07 and LRE09 for detecting American English in American English-Indian English language pair context and Hindi in Hindi-Urdu language pair context for each duration.

8. Language and Training Data Type

Since the training data supplied for each language was either all previously released conversational telephone speech (CTS) or newly released Voice of America (VOA) speech, it is of interest to compare performance by target language with separate plots for the 12 CTS training languages and for the 11 VOA training languages. This is presented for closed-set performance on three-second segments for one LRE09 system in Figure 12.

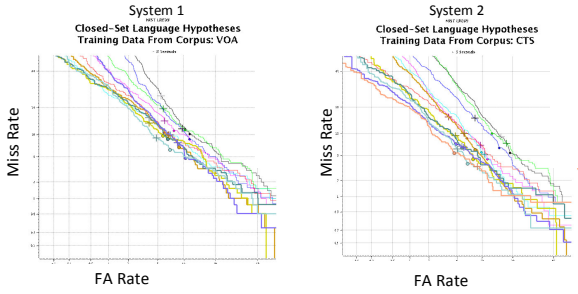


Figure 12: Closed-set performance on 3-second segments by target language for one LRE09 system. On left are the 12 languages with CTS training; on the right are the 11 languages with VOA training.

There is, unsurprisingly, a range of performance levels across languages based on various factors. There does appear to be a trend toward better performance on the CTS trained languages. This is probably due more to CTS trained languages having been tested in prior evaluations, unlike the VOA trained languages, than to properties of the training data itself. It may also be noted that the worst performing CTS trained languages were Indian languages, which may have particular confusability issues with one another.

9. Test Segment Data Type

It is also of interest to observe how performance was affected by the source (CTS or VOA) of the test segments. Figure 13 presents this comparison for closed-set results for each duration for three different LRE09 systems.

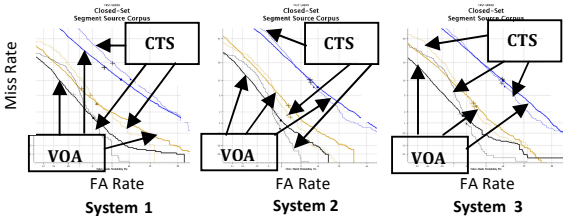


Figure 13: Closed-set performance by test segment source (CTS or VOA) for each duration for three LRE09 systems.

It is interesting to note that performance is broadly comparable on the two test data types. This is so even though for some of the VOA test languages the supplied training data was CTS. Also, as noted in the previous section, there was a slight trend toward better performance on languages with CTS training. It may also be noted that in Figure 13 the CTS curves appear less linear, tending to display relatively better performance at higher false alarm rates.

10. Summary and Future Plans

LRE09 represented an experiment involving a new paradigm for collecting data for this type of evaluation. The outcome was generally viewed as successful. There was widespread

participation by research sites from around the world. Overall performance on narrowband VOA data was found to be comparable with that on CTS data of the type used in prior evaluations. The use of VOA data allowed the inclusion of rather larger numbers of test segments per language and rather larger numbers of target and of out-of-set languages than in previous LREs in a cost effective manner. The auditing process appeared to work well for the selection of appropriate data, though there was concern that it did not avoid the inclusion of multiple segments from many individual speakers.

For the open and closed-set conditions there was evidence of continued performance improvement on the short 3-second segments compared with prior evaluations. This was gratifying, though there was concern that this could in part have been due to better auditing of such segments for speech content. Though performance on 30-second seconds degraded somewhat, the performance level here was already seen as very high.

The language-pairs condition, and particularly the emphasis on certain closely related language pairs of interest, showed that distinguishing certain pairs, particular ones that are mutually comprehensible, poses great performance challenges. (These pairs also pose considerable auditing challenges.)

There was considerable discussion at the evaluation workshop about the appropriateness of using DET curves to represent performance over multiple target languages, and this issue is likely to be revisited. Some considerations related to this are discussed in [10].

LRE09 was a successful experiment, but it remains to be seen if it will provide the model for further language recognition evaluation. Finding further existing large muliti-lingual corpora such as VOA will pose a considerable challenge itself. There is likely to be greater emphasis on the language-pair conditions, and on certain related language pairs of particular interest in future evaluations. But the timing and particulars of further ongoing evaluations of language recognition remain to be determined.

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

[2] NIST, Information Technology Laboratory, "The 2009 NIST Language Recognition Evaluation Results", *http://www.itl.nist.gov/iad/mig/tests/lre/2009/lre09_eval_results/index.html*

[3] NIST, Information Technology Laboratory, "Language Recognition Evaluation", *http://www.itl.nist.gov/iad/mig/tests/lre/2009/*

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

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

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

[7] Martin, A. et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. EUROSPEECH-97*, Rhodos, Greece, pp. 1895-1898, September 1997

[8] Wikipedia, "Hindi-Urdu Controversey", *http://en.wikipedia.org/wiki/Hindi-Urdu_controversy*

[9] Wikipedia, "Serbo_Croatian Language", *http://en.wikipedia.org/wiki/Serbo-Croatian*

[10] Brummer, N. and van Leeuwen, D, "On Calibration of Language Recognition Scores", *Proc. 2006 IEEE Odyssey – The Speaker and Language Recognition Workshop*, San Juan, PR, June 2006