*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# Results of The 2015 NIST Language Recognition Evaluation

*Hui Zhao[1*], Désiré Bansé[1*], George Doddington[1], Craig Greenberg[1], Jaime Hernández-Cordero[3], John Howard[1*], Lisa Mason[3], Alvin Martin[1], Douglas Reynolds[2], Elliot Singer[2], Audrey Tong[1]*

[1]National Institute of Standards and Technology, USA
[2]MIT Lincoln Laboratory, USA
[3]U.S. Department of Defense, USA
[*]Contractor

{audrey.tong|craig.greenberg|alvin.martin|desire.banse|hui.zhao|john.howard}@nist.gov
george.doddington@comcast.net
{dar|esinger}@ll.mit.edu,{jherna2|lpmicke}@tycho.ncsc.mil

## Abstract

In 2015, NIST conducted the most recent in an ongoing series of Language Recognition Evaluations (LRE) meant to foster research in language recognition. The 2015 Language Recognition Evaluation featured 20 target languages grouped into 6 language clusters. The evaluation was focused on distinguishing languages within each cluster, without disclosing which cluster a test language belongs to.

The 2015 evaluation introduced several new aspects, such as using limited and specified training data and a wider range of durations for test segments. Unlike in past LRE's, systems were not required to output hard decisions for each test language and test segment, instead systems were required to provide a vector of log likelihood ratios to indicate the likelihood a test segment matches a target language. A total of 24 research organizations participated in this four-month long evaluation and combined they submitted 167 systems to be evaluated. The evaluation results showed that top-performing systems exhibited similar performance and there were wide variations in performance based on language clusters and within cluster language pairs. Among the 6 clusters, the French cluster was the hardest to recognize, with near random performance, and the Slavic cluster was the easiest to recognize.

**Index Terms:** language recognition, language detection, language identification, NIST LRE, NIST evaluation

## 1. Introduction

The 2015 NIST Language Recognition Evaluation (LRE15) [1] was held in the autumn of 2015. It was the latest in a series of language recognition technology evaluations coordinated by NIST since 1996 [2]. Figure 1 shows the number of target languages and participants in all NIST LRE's. Recently, the number of target languages in LRE15 has declined slightly while the number of participants continues a rising trend.

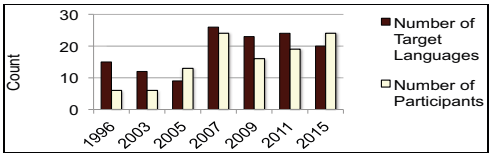



Figure 1: *Target language and participant statistics of the NIST LRE series.*

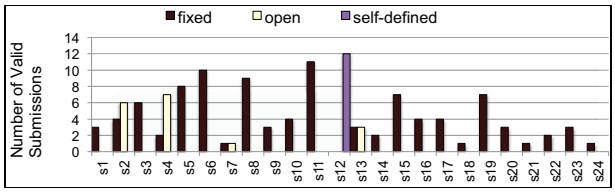



Figure 2. *Submission statistics of LRE15.*

The task in the NIST LRE's is language detection, i.e., given a test speech-recording and a set of target language speech-recordings, indicate whether the target language was spoken in the test speech-recording. This task in general was the focus task of NIST LRE's prior to 2011[2, 3]. Since LRE11 [4], the emphasis has shifted to distinguishing languages that are similar to each other and sometimes mutually intelligible. Similar to recent LRE's, LRE15 involved both conversational telephone speech (CTS) and broadcast narrowband speech (BNBS) data.

Unlike in past LRE's, there were two training conditions in LRE15: a fixed training condition and an open training condition. In the fixed training condition, participants could only use limited and specified training data to develop their systems and target language models. In the open training condition, additional data was permitted for use in system and model development. The fixed training condition was required of all LRE15 participants, and submissions to the open training condition were encouraged in order to help quantify the contribution of additional data on system performance. The number of submissions in LRE15 is shown in Figure 2. A total of 167 systems were submitted, of which 116 were verified by NIST to be valid, 99 for the fixed training condition and 17 for the open training condition.

In LRE15, test segments were selected to cover finer granularity of speech durations than prior LRE's. Instead of using recordings containing nominally 3, 10, or 30 seconds of speech, the LRE15 test segments were selected to have speech durations from the set of {3, 5, 10, 15, 20, 25, 30} seconds, which provided the opportunity to more precisely measure the effects of segment durations on performance.

Another key new aspect of LRE15 is that evaluated systems were not required to provide hard decisions for each target language / test segment pair. Instead, systems were required to provide a vector of real numbers, with entries



http://dx.doi.org/10.21437/Interspeech.2016-169

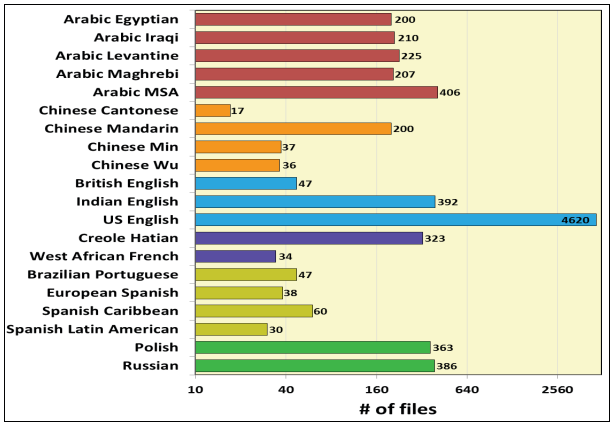



Figure 3. *Training set counts for each language in LRE15.*

interpreted as log likelihood ratios (llr) indicating the relative likelihood that a test segment was spoken in a target language rather than in another language in the same cluster as the target language.

## 2. Data

In LRE15, performance was evaluated by presenting systems with a series of speech recordings and target language speech recordings. In total there were twenty target languages grouped into six language clusters. While the evaluation focused on distinguishing languages within each cluster, the cluster to which each test speech recording belongs was not disclosed to the systems. Table 1 shows the target languages and language clusters in LRE15.

| Cluster | Target Languages |
|---|---|
| Arabic (ara) | Egyptian (arz), Iraqi (acm), Levantine (apc), Maghrebi (ary), Modern Standard (arb) |
| Chinese (chn) | Cantonese (yue), Mandarin (cmn), Min (cdo), Wu (wuu) |
| English (eng) | British (gbr), General American (usg), Indian (sas) |
| French (fre) | Haitian Creole (hat), West African (waf) |
| Slavic (sla) | Polish (pol), Russian (rus) |
| Iberian (spa) | Brazilian Portuguese (brz), Caribbean Spanish (car), European Spanish (eur), Latin American Spanish (lac) |

Table 1. *Target languages and language clusters in LRE15.*

Figure 3 shows the number of speech recordings available in the fixed training condition for each target language. Among all the languages, US English had the most recordings (4620), while Chinese Cantonese had the least (17). On average, there were 394 speech recordings per target language.

As described in the Introduction, the data used in LRE15 consisted of two collection types: broadcast narrowband speech (BNBS) and conversational telephone speech (CTS). Figure 4 depicts the collection type distribution of the target language training data. Training data for most languages (all except for Indian English, Brazilian Portuguese, Polish, and Russian) were drawn from a single collection type, which was predominantly CTS. This data consisted of speech recordings from several corpora collected by the Linguistic Data Consortium (LDC), namely CALLFRIEND[12], CALLHOME[12], LRE09/11 [3, 9], MIXER 3 [6],

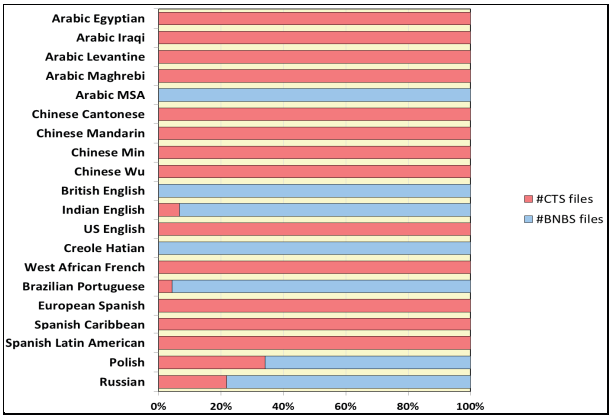



Figure 4. *Distribution of source type for training languages in LRE15.*

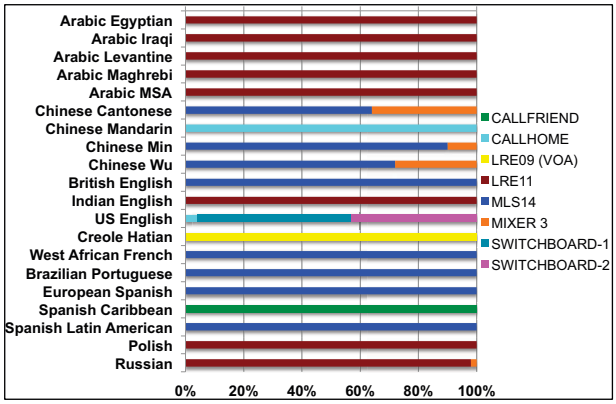



Figure 5. *Composition of training data from different corpora.*

SWITCHBOARD-1, and SWITCHBOARD-2 [13, 14], as well as a new corpus (MLS14) [10] collected specifically to support LRE15. Figure 5 shows the training data distribution by corpora for each language. The training data for eight of the twenty languages included recordings from the newly collected MLS14 corpus. Five languages had training data from at least two different corpora. The test speech recordings were drawn from two corpora: MLS14 and Babel, the latter of which was collected by Appen for the IARPA Babel program [7].

Figure 6 shows the collection type distribution of test speech recordings by language. Among the twenty languages, six were drawn from only one collection type, which, like the training data, was predominantly CTS. The majority of test segments for French-Haitian and Chinese-Cantonese languages were part of the Babel corpus.

Test speech recordings were created by NIST using an algorithm to select multiple recordings of varying duration from a single source-recording so as to minimize the overlap among the recordings. The distribution of speech duration for test recordings is shown in Figure 7. The majority of test recordings included in LRE15 were relatively short (less than 10 seconds). As a result, shorter recordings implicitly received larger weight in the computation of system performance.

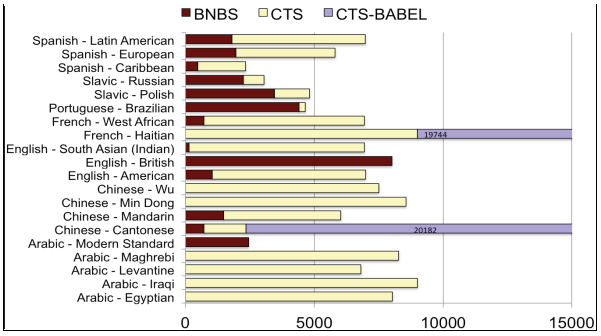



Figure 6. *Number of test segments by language source types.*

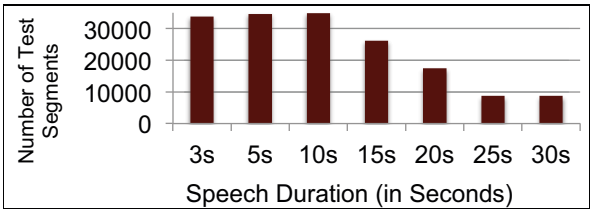



Figure 7. *Number of test segments by speech duration.*

## 3. Performance measurements

As described in the Introduction, systems submitted to LRE15 were required to provide a 20-dimentional vector of log likelihood ratios (llr) for each test segment. For each test speech recording and language L, a hard decision was inferred from its llr by comparing the llr with 0. An llr value of 0 in principle implies that, with equal priors and costs, it is as likely as not to be the language L. Thus if L were the true language for a test segment, a positive llr value would be treated as a correct detection, while a negative value would imply a miss.

Within each cluster, pair-wise language recognition performance was computed for all target/non-target language pairs (Lt, Ln). This was done in terms of detection miss and false alarm probabilities, and the miss and false alarm probabilities were computed separately for each target language and each target/non-target language pair, respectively. These probabilities were combined into a single number that represents the cost performance of a system, according to an application-motivated model:

$$C_{pair}(L_T, L_N) = C_{Miss} * P_{Target} * P_{Miss} + C_{FA} * \left(1 - P_{Target}\right) * P_{FA}(L_T, L_N) \quad (1)$$

where $L_T$ and $L_N$ are target language and non-target languages within a cluster, $P_{Miss}$ and $P_{FA}$ are the miss and false alarm probabilities, and $C_{Miss}$, $C_{FA}$ and $P_{Target}$ are application model parameters. In LRE15 $C_{Miss} = C_{FA} = 1$ and $P_{Target} = 0.5$.

The language pair costs were then averaged for each language cluster, as shown in equation (2):

$$C_{avg} = \frac{1}{N_L}\left\{\left[C_{Miss} * P_{Target} * \sum_{L_T} P_{Miss}(L_T)\right] + \frac{1}{N_L - 1}\left[C_{FA} * (1 - P_{Target}) * \sum_{L_T}\sum_{L_N} P_{FA}(L_T, L_N)\right]\right\} \quad (2)$$

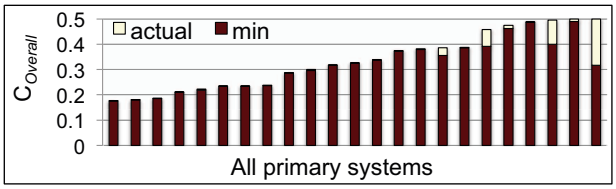
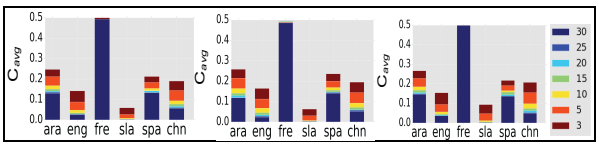



Figure 8. *Actual and minimum overall cost for primary submissions of LRE15.*




(a) system 1 (b) system 2 (c) system 3

Figure 9. *Results by language cluster for three top performing systems in LRE15, broken down by speech duration.*

where $N_L$ is the number of languages in the cluster.

It is worth noting that if a system always outputs positive valued vectors (i.e., a "no information" system), the system will have a $C_{avg}$ value of 0.5 (the same is true for a system that outputs only negative valued vectors). The basic $C_{avg}$ score for each cluster serves as primary performance measures for a system. In addition, the average of these values across the six clusters serve as a single overall performance cost for each system, called $C_{Overall}$.

$$C_{overall} = \frac{\sum_{N_C} C_{avg}}{N_C} \quad (3)$$

## 4. Results

Figure 8 shows the actual and minimum overall cost ($C_{Overall}$) for all of the primary submissions in the fixed training condition. Minimum cost is determined by varying the (system-wide) decision threshold from 0 so as to minimize the cost. We note that LRE15 system performance is difficult to compare with performance in prior LRE's due to difference in languages, amount of training data, test recording speech durations, and performance metrics.

The results per language cluster for three top-performing systems are shown in Figure 9, broken down by speech duration. Performance is very similar among these systems. More generally, performance on the Slavic language cluster tended to be the best, while performance on the French cluster tended to be the worst, with nearly random performance. These surprising results for the French cluster might be explained by training and test mismatch in Haitian Creole--the training data collection type was BNBS and the source collection type was CTS; the BNBS data tended to be a more formal variety of Haitian Creole (as might be used in news broadcasts) which is closer to West African French, and the CTS data tended to be less formal and more distinct from West African French. Further investigation of this result remains as future work.

Figure 10 shows the language pair costs for a top-performing system broken down by speech duration. We observe that costs vary widely for pairs within some language clusters. For example, for recordings with 30 seconds of

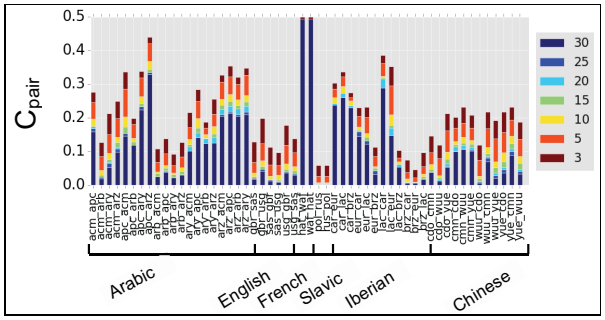



Figure 10. *Results by language pair Cpair(LT, LN) for a selected system. (Language names in each pair can be found in Table 1 through their abbreviations.)*

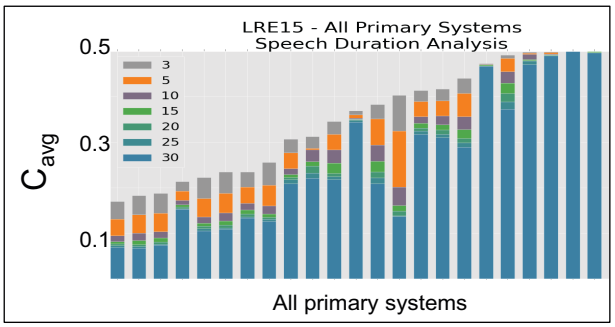



Figure 11. *Results by speech duration for participating systems.*

speech, the highest cost language pair in the Arabic cluster (ara) is about 16 times larger than the lowest cost language pair in the same cluster. Similarly, the highest cost is 9 times of the lowest cost in the cluster of Iberian (spa). We can also observe that even though Portuguese belongs to the Iberian cluster, it not confusable with other Iberian varieties. However, some languages in the same cluster are easily confusable with each other, such as Egyptian and Levantine Arabic as well as Latin American and Caribbean Spanish. It is interesting to note that language pair performance is not symmetric. For example, $C_{Pair}$(ara-arz,ara-arb) is almost 7 times $C_{Pair}$(ara-arb, ara-arz).

In Figure 11 we see performance for all primary systems broken down by speech duration. There is limited performance difference for speech durations between 20 seconds and 30 seconds. However, there is sharp drop in performance when the speech duration changes from 10 seconds to 5 seconds and similarly from 5 seconds to 3 seconds. This indicates that when the test recording speech duration is relatively short (below 10 seconds), additional speech in the test recording sharply improves performance (and this is less true when there are at least 10 seconds of speech).

Figure 12(a) shows the results by training condition for three top performing systems. We observe limited improvement in the open training condition over the fixed training condition (and, in one case, worse performance, which the participant attributes to using training data in the open training condition that was too mismatched from the test data).

Figure 12 (b) and Figure 12 (c) show the results based on gender and test segment collection type. Little performance difference was observed between male and female speakers. System performance on CTS speech recording is somewhat

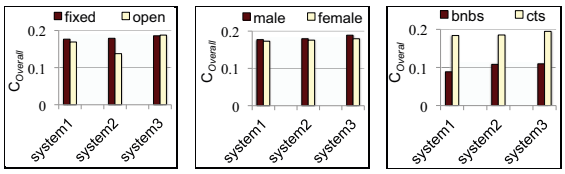



Figure 12. *Results by selected language characteristics for the top three systems. (a) Results by training condition. (b) Results by speaker gender. (c) Results by source type.*

worse than on BNBS, and we believe this is due to which languages had predominantly CTS test segments.

## 5. Conclusion and Future Work

We present a summary of the 2015 NIST Language Recognition Evaluation. The objective of LRE15 was to provide a platform for evaluating the most advanced technology in language recognition and to foster new ideas and collaboration. LRE15 attracted worldwide participants and had the largest number of participants in its history.

The biggest change in LRE15 was the inclusion of a fixed training condition. This new aspect is likely to be refined and improved upon in future LREs. The nearly random performance on the French cluster was surprising and further exploration of this result remains interesting future work. We observed that for some language clusters, the top systems have significantly better performance than the rest of the systems. Another insight from this evaluation is that more training data does not lead to better performance if the data is not used properly. Additional analysis of system performance results is also planned for future evaluations.

LRE15 was deemed a success, and there are plans for a follow on analysis workshop, to be held in late 2016, as well as a new LRE, to be held during 2017.

## 6. 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.

## 7. References

[1] NIST Language Recognition Evaluation. Available: http://www.nist.gov/itl/iad/mig/lre.cfm

[2] A. F. Martin and et al., "NIST Language Recognition Evaluation – Past and Future," in *Proceedings Odyssey*, Joensuu, Finland, 2014, pp. 145-151.

[3] A. F. Martin and C. S. Greenberg, "The 2009 NIST Language Recognition Evaluation", in *Proceedings Odyssey*, Brno, Caech Republic, 2010.

[4] C. S. Greenberg, A. F. Martin and M. A. Przybocki, "The 2011 NIST Language Recognition Evaluation", in *INTERSPEECH*, Portland, USA, 2012.

[5] M. Liberman and C. Cieri, "The Creation, Distribution and Use of Linguistic Data," in *LREC 1998*, Granada, Spain, 1998.
[6] C. Cieri, and et al., "The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data," in *LREC 2004*, Lisbon, Portugal, 2004, pp. 627-630.
[7] M. P. Harper, "Data Resources to Support the Babel Program Intelligence Advanced Research Projects Activity (IARPA)," [Online], Available: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/harper.pdf
[8] C. Dieri, L. Brandschain, A. Neely, D. Graff, K. Walker, C. Caruso, A. F. Martin and C. S. Greenberg, "The Broadcast Narrow Band Speech Corpus: a New Resource Type for Large Scale Language Recognition," In *INTERSPEECH*, Brighton, England, 2009.
[9] S. Strassel, and et al., "New Resources for Recognition of Confusable Linguistic Varieties: The LRE11 Corpus," in *Proceedings Odyssey*, Singapore, 2012, pp. 202-208.
[10] LDC, "LDC 2014 Multi-Language Speech Collection Corpus Specification", 2014.
[11] A. Martin, C. Greenberg, D. Graff, K. Walker and L. Brandschain, "2009 NIST Language Recognition Evaluation Test Set", [Online], Available: https://catalog.ldc.upenn.edu/LDC2014S06
[12] J. Benesty, M. M. Sondhi, Y. Huang "Springer Handbook of Speech Processing", Springer Science & Business Media, Nov 28, 2007.
[13] LDC, "Switchboard-1 Release 2", [Online], Available: https://catalog.ldc.upenn.edu/LDC97S62
[14] LDC, "Switchboard-2 Phase II", [Online], Available: https://catalog.ldc.upenn.edu/LDC99S79