The 2022 NIST Language Recognition Evaluation

Yooyoung Lee[1], Craig Greenberg[1], Eliot Godard[1,], Asad A. Butt[1,*], Elliot Singer[2,+],*
Trang Nguyen[2,+], Lisa Mason[3], Douglas Reynolds[3]

[1]NIST ITL/IAD/Multimodal Information Group, MD, USA
[2]MIT Lincoln Laboratory, Lexington, MA, USA
[3]U.S. Department of Defense, MD, USA

lre_poc@nist.gov

Abstract

In 2022, the U.S. National Institute of Standards and Technology (NIST) conducted the latest Language Recognition Evaluation (LRE) in an ongoing series administered by NIST since 1996 to foster research in language recognition and to measure state-of-the-art technology. Similar to previous LREs, LRE22 focused on conversational telephone speech (CTS) and broadcast narrowband speech (BNBS) data. LRE22 also introduced new evaluation features, such as an emphasis on African languages, including low resource languages, and a test set consisting of segments containing between 3s and 35s of speech randomly sampled and extracted from longer recordings. A total of 21 research organizations, forming 16 teams, participated in this 3-month long evaluation and made a total of 65 valid system submissions to be evaluated. This paper presents an overview of LRE22 and an analysis of system performance over different evaluation conditions. The evaluation results suggest that Oromo and Tigrinya are easier to detect while Xhosa and Zulu are more challenging. A greater confusability is seen for some language pairs. When speech duration increased, system performance significantly increased up to a certain duration, and then a diminishing return on system performance is observed afterward.

Index Terms: human language technology, LRE, language recognition, language detection, speech technology performance evaluation

1. Introduction

The 2022 NIST Language Recognition Evaluation (LRE), held in fall of 2022, was the latest in an ongoing series of language recognition evaluations conducted by NIST since 1996 [1]. The primary objectives of the LRE series are to: 1) advance language recognition technologies with innovative ideas, 2) facilitate the development of language recognition technology by providing data and research direction, and 3) measure the performance of the current state-of-the-art technology. Figure 1 shows the number of target languages and participants (based on sites) for all NIST LREs.

LRE22 was conducted entirely online using a web-based platform like LRE15 [2] and LRE17 [3, 4]. The updated LRE22 web-platform[1] supported a variety of evaluation activities, such
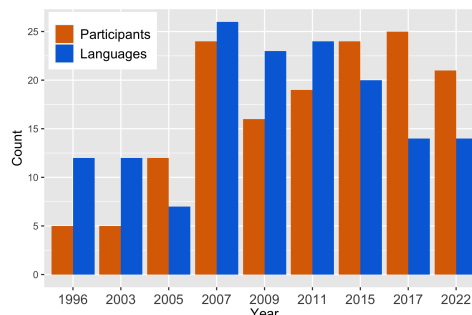
Figure 1: *Language and participant count for the NIST LREs*
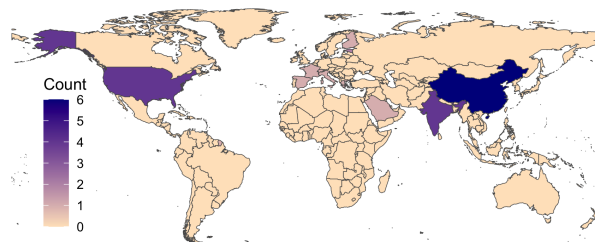

Figure 2: *Heatmap of the world showing the number of LRE22 participating sites per country.*

as registration, data license submission, data distribution, system output submission and validation/scoring, and system description/presentation uploads. A total of 16 teams from 21 organizations in 13 different countries made submissions for LRE22. Figure 2 displays a world map with heatmap representing the number of participating sites per country. Since two teams did not submit valid system descriptions, analysis considering only 14 teams in presented this paper. It should be noted that all participant information, including country, was self-reported.

2. Task

The general task in the NIST LREs is language detection, i.e. to automatically determine whether a particular target language was spoken in a given test segment of speech. Since LRE11 [5], the focus of the language detection task had turned to distinguishing between closely related, and sometimes mutually intelligible, languages. However LRE22 introduced a new emphasis on distinguishing between African languages, including low resource languages. Table 1 shows the 14 target languages included in LRE22. Similar to LRE17, LRE22 participants were required to provide a 14-dimensional vector of log-likelihood scores corresponding to the languages in Table 1.

10.21437/Interspeech.2023-241

Table 1: *LRE22 target languages*

Language	Code	Language	Code
Afrikaans	afr-afr	Ndebele	nbl-nbl
Tunisian Arabic	ara-aeb	Oromo	orm-orm
Algerian Arabic	ara-arq	Tigrinya	tir-tir
Libyan Arabic	ara-ayl	Tsonga	tso-tso
South African English	eng-ens	Venda	ven-ven
Indian-accent South African English	eng-iaf	Xhosa	xho-xho
North African French	fra-ntf	Zulu	zul-zul

Unlike LRE17, language clusters were not considered in this evaluation.

Like LRE17, there were two training conditions in LRE22: *fixed* and *open*. For the *fixed* training condition, participants were restricted to use only a limited pre-specified set of data for system training and target model development. For the *open* training condition, participants were allowed to utilize unlimited amounts of publicly available and/or proprietary data for their system training and target model development. To facilitate more meaningful cross-system comparisons, LRE22 participants were required to provide submissions to the *fixed* condition while participation in the optional *open* condition was strongly encouraged to understand the impacts that larger amounts of training and development data have on system performance. In order to encourage participation in the *open* training condition, the deadline for this condition was made one week later than the required *fixed* training condition submission deadline. A total of 65 valid submissions were received, 40 for the *fixed* training condition and 25 for the *open* condition. LRE participants were required to specify one submission as *primary* for each training condition they took part in, while all other systems submitted were considered *alternate*.

3. Data

This section provides a brief description of data used in LRE22 for training, development (*dev*), and evaluation (*test*) sets, along with the associated metadata.

3.1. Training set

As mentioned in Section 2, there were two training conditions in LRE22. The *fixed* condition limited the system training and development data to the following specific data sets provided to participants by the Linguistic Data Consortium (LDC): 2017 NIST LRE *dev* set and previous NIST LRE training data (LDC2022E16), 2017 NIST LRE *test* set (LDC2022E17), 2022 NIST LRE *dev* set (LDC2022E14). The VoxLingua107 data set [6] was also permitted for use in the *fixed* condition. The *open* training condition removed the limitations of the *fixed* condition. In addition to the data listed in the *fixed* condition, participants could use any additional data to train and develop their system, including proprietary data and data that are not publicly available. LDC also made selected data from the IARPA Babel Program [7] available to participants to be used in the *open* training condition.

3.2. Development and test sets

The development (*dev*) set is normally used to build/optimize a system model during the development process while the evaluation (*test*) set is used to evaluate the performance of the system model. The speech segments in the LRE22 *dev* and *test*

sets were selected from data sets collected by the Linguistic Data Consortium (LDC) to support LR technology evaluations; namely the Maghrebi Language identification (MAGLIC) [8], Speech Archive of South African Languages (SASAL) [9], and Low Resource African Languages (LRAL) [9] corpora. The MAGLIC corpus was a CTS-only collection based in Tunisia and includes four regional language varieties spoken in North Africa: Algerian Arabic, Libyan Arabic, Tunisian Arabic, and North African French. The SASAL corpus was a conversational telephone speech(CTS) and broadcast narrowband speech (BNBS) collection located in South Africa and contains several African language varieties, a subset of which were included in LRE22: Afrikaans, Ndebele, Tsonga, Venda, Xhosa, and Zulu, as well as South African English and Indian-accented South African English. The LRAL corpus was a BNBS collection based in Ethiopia, and, of the languages in LRAL, two were selected for inclusion in LRE22: Oromo and Tigrinya.
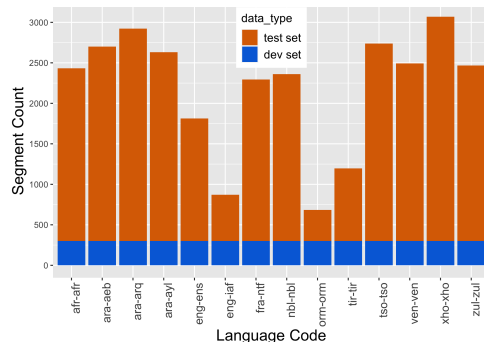

Figure 3: *Distribution of speech segments per target language for both dev and test sets*

All audio data provided was sampled at 8 kHz, a-law encoded, and formatted as SPHERE [10] files. When the source audio recordings were higher bandwidth or encoded differently, they were downsampled and transcoded to 8-kHz a-law. Unlike in previous LREs, the amount of speech in the LRE22 segments was uniformly sampled between approximately 3 and 35 seconds, as determined by an automatic speech activity detector. Figure 3 shows a stacked histogram for the *dev* and *test* sets. The *dev* set consisted of 300 segments per target language while the *test* set contained a total of 26,473 segments ranging from 383 to 2,769 segments across the target languages.

3.3. Metadata

The metadata collected by LDC can be categorized into audio- and audit-related metadata. The audio metadata indicates information related to the audio recording or segment, such as speech duration, data source type (i.e., either CTS or BNBS), and source file (i.e., the original recording from which the audio segment was extracted). The audit metadata reflects a human auditor's judgement of the speech, having listened to an audio recording, such as whether the recording contained a single speaker, if the person speaking was a native speaker, the speech clarity, the speaker sex, or if the recording took place in a noisy environment. In this paper, we limit our analyses on data source type and speech duration.

4. Performance Measure

As stated in the Section 2, LRE22 participants were required to provide a 14-dimensional vector of log-likelihood scores
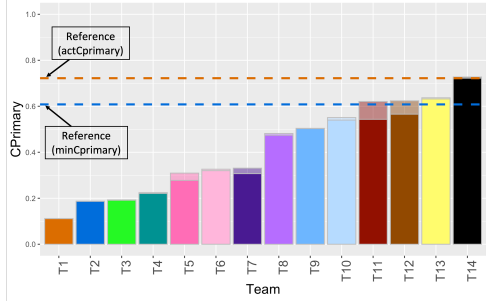
Figure 4: *System performance (actual and minimum costs) on primary submissions under the fixed training condition*
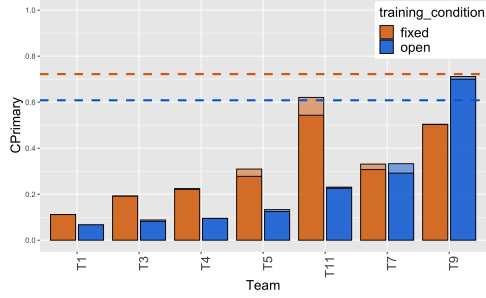

Figure 5: *A performance comparison of the fixed and open training conditions*

for the 14 target languages (see Table 1 for the LRE22 target languages). Pair-wise performance was computed for all target/non-target language pairs. A decision threshold derived from log-likelihood ratios was used to determine the number of missed detections and false alarms, computed separately for each target language. The missed detections (Misses) indicate the segments that are the target language, but are not predicted to be, while the false alarms (FAs) indicate the segments that are falsely identified as the target language. The probabilities of missed detections (P_{Miss}) and false alarms (P_{FA}) are then combined using a linear cost function [11]:

$$C(L_T, L_N) = C_{Miss} \times P_{Target} \times P_{Miss}(L_T) +$$
$$C_{FA} \times (1 - P_{Target}) \times P_{FA}(L_T, L_N) \quad (1)$$

where L_T and L_N are target and non-target languages, respectively. Here, C_{Miss} (cost of a missed detection), C_{FA} (cost of a false alarm), and P_{Target} (the *a priori* probability of the specified target language) are application-motivated cost model parameters. Two sets of cost-function parameters were used in LRE22: the first set of parameters provides equal weighting to the costs of errors ($C_{Miss} = C_{FA} = 1$) and a target probability of 0.5, while the second set of parameters changed the target probability to 0.1. The final metric, $C_{Primary}$, consisted of the mean value of the costs using the two different cost function parameters, normalized by dividing by the cost of a "no information" system. Costs using thresholds that minimize the Bayes risk, $actC_{Primary}$, as well as using thresholds that minimize the empirical cost, $minC_{Primary}$, were computed. We refer readers to the LRE22 evaluation plan [11] for details of the performance measures.

5. Results and Analyses

A total of 14 teams from academic and industrial sectors successfully completed LRE22. For both the *fixed* and *open* training conditions, the teams were allowed to have one *primary* sub-

mission and one or more *alternate* submissions. In this section, we present a summary of results and key findings on the *primary* submissions using the metrics defined in Section 4.

Figure 4 illustrates system performance for all the *primary* submissions under the *fixed* training condition. The x-axis are anonymized team names and the y-axis are $C_{Primary}$ values for both the actual and minimum costs (N.B., a lower $C_{Primary}$ value indicates better performance). The orange dashed-line indicates an actual cost, $actC_{Primary}$, and the blue is a minimum cost, $minC_{Primary}$, for a reference system; we used an off-the-shelf algorithm as a reference to validate the LRE22 data construction and evaluation process. The reference system was trained and fine-tuned only on VoxLingua107 and the LRE22 development set. The shaded color on each team's bar indicates the difference between $actC_{Primary}$ and $minC_{Primary}$, which indicicates a calibration error. In Figure 4, we observe that, given the primary submissions under the *fixed* condition, the $C_{Primary}$ values range from 0.11 to 0.73 across all the teams. It is observed that the top-performing systems (e.g., T1-T4) have small calibration errors (i.e., the absolute difference between the actual and minimum costs is relatively small) while a few teams (e.g., T5, T7, T11 and T12) are less well-calibrated.
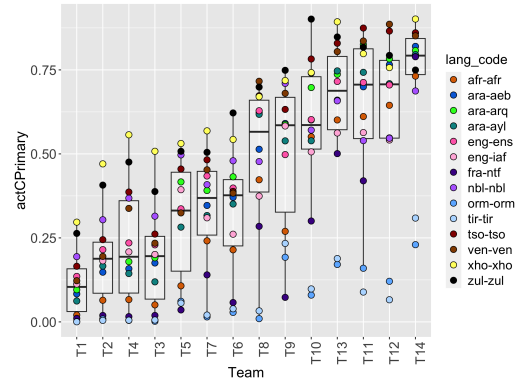

Figure 6: *A language-level performance on primary submissions under the fixed training condition*

As described in Section 2, the *fixed* training condition is required while *open* is optional; 7 out of the 14 teams submitted their system outputs to the *open* training condition. Figure 5 illustrates a performance comparison of training conditions (*fixed* vs *open*) for the seven teams only (ordered by *open* system performance). The result shows that system performance from the *open* condition generally outperforms the *fixed* condition submissions across the teams (except T9), and a calibration error is observed in team T7 under the *open* training condition.

To understand variability of language-level system performance and language detection difficulty, Figure 6 illustrates a box plot of the primary submission performance under the *fixed* training condition. The x-axis is a team name (ordered by median), the y-axis is the actual cost ($actC_{Primary}$), and each point represents a target language. The black line within a box is the median, the box edges represent the lower quartile and upper quartile, and the whiskers extending from the box indicate variability outside the upper and lower quartiles. We observe a high dispersion of language performance for a few teams such as T4, T5, and T9. Overall, the *Oromo (orm-orm)* and *Tigrinya (tir-tir)* points marked in blue are located in the bottom side of Figure 6 (easier to detect) while *Xhosa (xho-xho)* and *Zulu (zul-zul)* are in the top (harder to detect); a similar trend is observed across the teams.
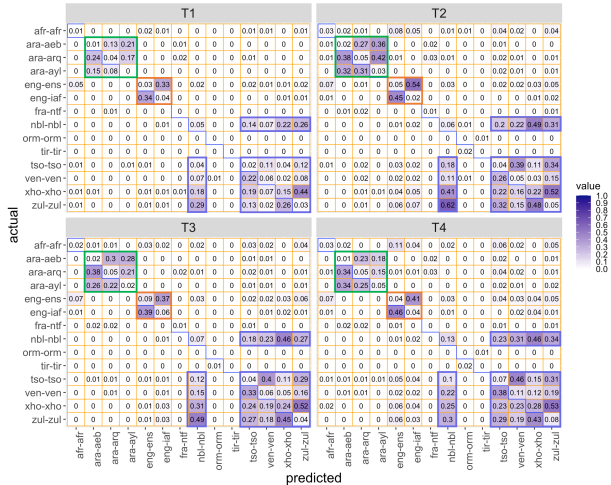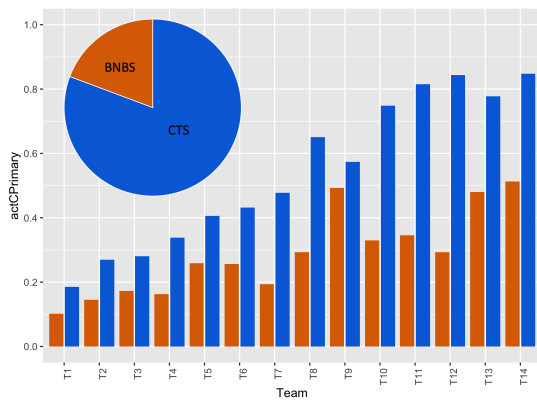
Figure 7: *Language confusability of the leading systems*

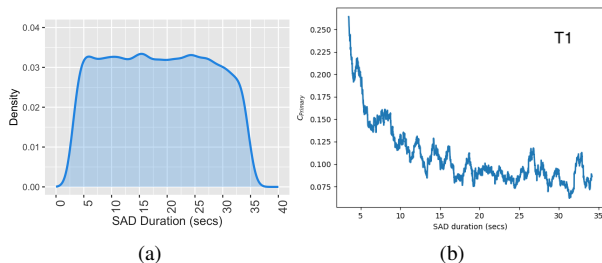Figure 8: *A data source type distribution and effect on systems*

Figure 9: *SAD duration effect on system performance (a) SAD duration distribution for the dev and test set (b) T1 system performance vs. SAD duration.*

To examine language-pair confusability, we conducted data analysis using heatmap confusion matrices as shown in Figure 7. The axes are language codes. The diagonal values from upper-left to bottom-right are P_{Miss} (false reject rates) and the off-diagonal values are P_{FA} (false alarm rates). A higher false alarm probability implies a potential confusability for that language pair. For simplicity, results of $P_{Target} = 0.5$ for the four leading systems are demonstrated using heatmap confusion matrices. Given the *test* set and systems, a higher confusability is observed for three clusters of language pairs as follows: 1) among Arabic languages (ara-aeb, ara-arq, ara-ayl), 2) between South African English (eng-ens) and Indian-accented South African English (eng-iaf), and 3) Ndebele (nbl-nbl), Tsonga

(tso-tso), Venda (ven-ven), Xhosa (xho-xho) and Zulu (zul-zul). According to the references [9][12], Zulu, Xhosa, Ndebele, Tsonga, and Venda are Bantu languages and have some mutual intelligibility but can be easily distinguished from one another by native speakers.

To gain insight on how metadata variables (i.e., factors) affect system performance, we conducted experiments given the metadata listed in Section 3.3. For simplicity, the following analyses are demonstrated using *data source type* and *speech duration* only. The LRE22 data was collected in two primary genres, CTS and BNBS, which we call *data source type*. Figure 8 shows system performance ($actC_{Primary}$) partitioned by *data source type* (CTS vs BNBS) for all the *primary* submissions under the *fixed* training condition. The top-left pie chart is a distribution of CTS and BNBS on the *test* set, which is imbalanced. The bar plot shows a performance comparison between CTS (blue) and BNBS (orange) across all the teams. The results indicate that, given the imbalanced distribution, CTS is more challenging and that *data source type* has a strong effect on system performance; a similar trend is observed across the systems.

Durations of *test* set segments varied between 3s and 35s of speech that have been randomly sampled and extracted from longer recordings as determined by an automatic Speech Activity Detector (SAD) which we call *SAD duration*. Figure 9a shows a distribution of *SAD duration* for the *test* set and Figures 9b shows the performance of a top-performing system by *SAD duration*. Given the *test* set and systems, it is seen that when *SAD duration* increases, $actC_{Primary}$ significantly decreases up to a certain duration (between 15s and 20s). After that, a diminishing return on system performance improvement is observed across the systems.

6. Conclusions

We presented a summary of the 2022 NIST Language Recognition Evaluation with an emphasis on low resource languages and random duration of speech segments.

The results showed that almost no calibration error was observed for the top-performing systems for both the *fixed* and *open* training condition. Overall, the submissions under the *open* training condition had better performance compared to the *fixed* condition submissions, with only one exception. Given the *test* set and *primary* systems under the *fixed* training condition, we found that Oromo and Tigrinya were easier to detect while Xhosa and Zulu were harder to detect. A greater confusability was observed for the language pairs 1) among Zulu, Xhosa, Ndebele, Tsonga, and Venda, 2) between South African and Indian-accent South African English, and 3) among Tunisian, Algerian, and Libyan Arabic languages. Some of the metadata, such as *data source type* and *SAD duration*, had a significant effect on system performance for all systems. In terms of *SAD duration*, when speech duration increased, system performance significantly increased up to a certain duration, and then we observed a diminishing return on system performance afterward.

7. Disclaimer

These results presented in this paper 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.

8. References

[1] National Institute of Standards and Technology, "NIST Language Recognition Evaluation Overview," https://www.nist.gov/itl/iad/mig/language-recognition, 1996-2022.

[2] H. Zhao, D. Bansé, G. Doddington, C. Greenberg, J. Hernández-Cordero, J. Howard, L. Mason, A. Martin, D. Reynolds, E. Singer, and A. Tong, "Results of The 2015 NIST Language Recognition Evaluation," in *Interspeech 2016*, San Francisco, USA, September 2016, pp. 3206–3210.

[3] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and Hernandez-Cordero, "The 2017 NIST Language Recognition Evaluation," in *Odyssey*, 2018, pp. 82–89.

[4] S. O. Sadjadi, T. Kheyrkhah, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "Performance Analysis of the 2017 NIST Language Recognition Evaluation," in *Proc. Interspeech 2018*, 2018, pp. 1798–1802.

[5] A. F. Martin, C. S. Greenberg, J. M. Howard, G. R. Doddington, and J. J. Godfrey, "NIST Language Recognition Evaluation - Past and Future," in *Odyssey 2014*, Joensuu, Finland, June 2014, pp. 145–151.

[6] J. Valk and T. Alumäe, "Voxlingua107: A dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 652–658.

[7] M. P. Harper, "Data Resources to Support the Babel Program," https://goo.gl/9aq958, 2011.

[8] K. Jones, K. Walker, C. Caruso, S. Strassel, and J. Wright, "MAGLIC: Maghrebi Language Identification Corpus," in *Inter-Speech submitted*, 2023, p. na.

[9] K. Jones, C. Caruso, K. Walker, and S. Strassel, "LDC Data for LRE2," 31 January 2023. [Online]. Available: https://www.nist.gov/news-events/events/2023/01/2022-language-recognition-evaluation-lre-workshop

[10] National Institute of Standards and Technology, "Speech File Manipulation Software (SPHERE) Package Version 2.7," ftp://jaguar.ncsl.nist.gov/pub/sphere-2.7-20120312-1513.tar.bz2, 2012.

[11] Y. Lee, C. Greenberg, L. Mason, and E. Singer, "The 2022 NIST Language Recognition Evaluation Plan," https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=935161, 2022.

[12] M. Alexander, "The 11 languages of south africa," 6 March 2018, [Online; accessed 1-May-2023]. [Online]. Available: https://southafrica-info.com/arts-culture/11-languages-south-africa