NIST 2022 Language Recognition Evaluation Plan

August 31, 2022

1 Introduction

The 2022 NIST language recognition evaluation (LRE22) is the 9$^{\text{th}}$ cycle in an on-going language recognition evaluation series that began in 1996. The objectives of the evaluation series are (1) to advance technologies in language recognition with innovative ideas, (2) to facilitate the development of language recognition technology incorporating these ideas, and (3) to measure the performance of the current state-of-the-art technology. Targeting researchers working on the general problem of text-independent, speaker-independent language recognition, the evaluation is designed to focus on core technology issues and to be simple and accessible to those wishing to participate.

LRE22 will be organized in a similar manner to LRE17 [1], focusing on conversational telephone speech (CTS) and broadcast narrow band speech (BNBS) data. As in LRE17, *fixed* and *open* training conditions will be offered to allow cross-system comparisons and to understand the effect of additional and unconstrained amounts of training data on system performance.

A new focus in LRE22 is an emphasis on language recognition of African languages, including low-resource languages. An additional new feature for LRE22 is expanded test segment durations. In prior LREs, test segments from CTS and BNBS were extracted from longer recordings to create smaller chunks containing approximately 3s, 10s, or 30s of speech. The LRE22 evaluation data will consist of segments that contain between 3s and 30s of speech that have been randomly sampled and extracted from longer recordings. NIST will distribute a small development set representative of the test set to participants. Participants will be asked to provide score vectors with the log-likelihood scores, rather than the log-likelihood ratios.

Participation in LRE22 is open to all who find the evaluation of interest and are able to comply with the evaluation rules set forth in this plan. There is no cost to participate, but participating teams must be represented at the LRE virtual workshop planned for January 2023. Information about evaluation registration can be found on the LRE22 website[1].

2 Task Description

2.1 Task Definition

The task for LRE22 is *language detection*: given a segment of speech and a target language, automatically determine if the target language was spoken in the test segment. LRE22 has 14 target languages listed in Table 1. Unlike LRE17, language clusters will not be defined and will not be considered in the LRE22 primary evaluation metric.

Input to a language recognition (LR) system will be a series of test segments, and the output from the system will be a series of score vectors, one vector per test segment. Each score vector is defined as a 14-dimensional vector corresponding to the 14 target languages in the order listed in Section 6.4, and representing estimated log-likelihood scores, using natural (base e) logarithms, for the corresponding languages.

[1] https://lre.nist.gov

In terms of the conditional probabilities for the observed data (O) given a target language model (L_i), the log-likelihood score (ℓ_i) is defined as

$$\ell_i = \log(P(O|L_i)). \tag{1}$$

The likelihood function in (1) is related to the posterior probability $P(L_i|O)$ via Bayes' rule as follows

$$P(L_i|O) = \frac{P(L_i)\exp(\ell_i)}{\sum_{j=1}^{N_L} P(L_j)\exp(\ell_j)}, \tag{2}$$

where $P(L_i)$ is the *a priori* probability of the language class i, and N_L is the number of target languages.

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

Table 1: LRE22 target languages

2.2 Training Conditions

The training condition is defined as the amount of data/resources used to build an LR system. The task described above can be evaluated over a *fixed* (required) or *open* (optional) training condition.

- **Fixed** – The *fixed* training condition limits the system training to the following specific data sets:
  - 2017 NIST LRE Development Set and previous[2] NIST LRE training data (LDC2022E16)
  - 2017 NIST LRE Test Set (LDC2022E17)
  - 2022 NIST LRE Development Set (LDC2022E14)

  Participants can obtain the data from the LDC after they have signed the data license agreement. In addition to these data, participants may also use the VoxLingua107 corpus[3] [2].

  For the *fixed* training condition, only the specified speech data may be used for system training and development, including all system modules (e.g., speech activity detection) and auxiliary systems employed to build an LR system (e.g., automatic speech recognition). Publicly available non-speech audio and data (e.g., noise samples, impulse responses, filters) may be used and must be noted in the system description. Participation in the *fixed* condition is required.

[2]as released in LRE17
[3]http://bark.phon.ioc.ee/voxlingua107/

Note: The use of pretrained models on data other than what is designated above (e.g., BUT Hungarian phoneme recognizer) is not allowed in this condition.

- **Open –** The *open* training condition removes the limitations of the *fixed* condition. In addition to the data listed in the *fixed* condition, participants can use any additional data including proprietary data and data that are not publicly available. Please note that any additional data used must be adequately described by providing enough details in the system description.

  LDC will also make available selected data from the IARPA Babel Program to be used in the *open* training condition. Participation in this condition is optional but *strongly* encouraged to demonstrate the gains that can be achieved with unconstrained amounts of data.

3 Performance Measurement

3.1 Primary Metric

Pair-wise LR performance is computed for all target/non-target language pair scores (L_T, L_N). A decision threshold is used to determine counts of missed detection (Miss) and false alarms (FA) computed separately for each language. The missed detections (false rejects) indicate the segments that are falsely predicted as a non-target language, while the false alarms (false accepts) indicate the segments that are falsely identified as a target language. The probability of missed detection (P_{Miss}) and false alarms (P_{FA}) are then combined using a linear cost function according to an application-motivated cost model, defined as

$$\begin{aligned} C(L_T, L_N) = & C_{Miss} \times P_{Target} \times P_{Miss}(L_T) + \\ & C_{FA} \times (1 - P_{Target}) \times P_{FA}(L_T, L_N), \end{aligned} \tag{3}$$

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 model parameters and defined to have the following values:

Parameter ID	C_{Miss}	C_{FA}	P_{Target}
1	1	1	0.5
2	1	1	0.1

Table 2: LRE22 cost parameters

The first set of parameters provides equal weighting to the costs of errors (missed detections and false alarms) and the target probability, while the second set of parameters is changed so that the target probability is 0.1. To improve the interpretability of the cost function, it will be normalized by $C_{Default}$, which is defined as the best cost that could be obtained without processing the input data (i.e., by either always accepting or always rejecting the segment language as matching the target language, whichever gives the lower cost) as follows

$$C_{Norm}(L_T, L_N) = C(L_T, L_N)/C_{Default}, \tag{4}$$

Here, the default cost for both sets of parameters defined in Table 2 is set to $C_{Default} = C_{Miss} \times P_{Target}$. Rewriting the cost model in (3) by combining all of the application model parameters yields

$$C_{Norm}(L_T, L_N) = P_{Miss}(L_T) + \beta \times P_{FA}(L_T, L_N), \tag{5}$$

where β is defined as:

$$\beta = \frac{C_{FA} \times (1 - P_{Target})}{C_{Miss} \times P_{Target}}.$$

Actual detection costs will be computed by applying detection thresholds of $\log(\beta)$ to log-likelihood *ratios* derived from the log-likelihoods output by the system [4].

In addition to the performance measures computed for each target/non-target language pair, an average cost performance for each system will be computed as

$$C_{avg}(\beta) = \frac{1}{N_L}\left\{\sum_{L_T} P_{Miss}(L_T) + \frac{1}{N_L - 1}\left[\beta \times \sum_{L_T}\sum_{L_N} P_{FA}(L_T, L_N)\right]\right\}, \tag{6}$$

where N_L is the number of target languages. The primary metric for LRE22 will be the average cost performance defined in (6), computed using the two application model parameters given in Table 2, that are then averaged:

$$C_{primary} = \frac{C_{avg}(\beta_1) + C_{avg}(\beta_2)}{2}. \tag{7}$$

NIST will make available the script that calculates the primary metric.

3.2 Secondary Metric

In addition to the cost metric C_{avg} described above, an alternative information theoretic metrics may also be used to calculate the performance of an LR system at the discretion of evaluators and participants. For instance, the multiclass cross-entropy metric H_{mce} measures the information the LR system provides through the log-likelihood scores and is defined as follows [3]:

$$H_{mce} = -\sum_{i=1}^{N_L} \frac{P(L_i)}{\|S_i\|} \sum_{t \in S_i} \log P(L_i|O_t), \tag{8}$$

where S_i is the subset of indices for segments of target language i, $\|S_i\|$ is the number of segments of target language i.

For a *do-nothing* default system, the multiclass cross-entropy is given by

$$H_{max} = -\sum_{i=1}^{N_L} P(L_i) \log P(L_i). \tag{9}$$

If $H_{mce} \geq H_{max}$ for an LR system, then it does not improve upon the default *do-nothing* system. To facilitate the interpretation of the cross-entropy or mutual information, a normalized version of H_{mce} is calculated as *confidence* score which is defined as [4]:

$$Confidence = 1 - \frac{H_{mce}}{H_{max}}. \tag{10}$$

Given that the cross-entropy is non-negative, a perfect LR system achieves a confidence score of 1 (i.e., it has zero confusion), while a totally confused system can achieve a confidence score of zero (or less).

[4] Log-likelihood ratios will be computed as the difference between the target language log-likelihood and the linear average of the log-likelihoods of the non-target languages, i.e., $LLR(L_i) = -\log\left[\frac{1}{N_L - 1}\sum_{j \neq i} \exp\left(\ell_j - \ell_i\right)\right]$.

4 Development and Test Data Description

New data collected outside of North America by the LDC from the following corpora will be used to compile the LRE22 development and test sets: the Maghrebi Language Identification Corpus (MAGLIC), the Speech Archive of South African Languages (SASAL) corpus, and the Low Resource African Languages (LRAL) corpus. A small subset extracted from these corpora will also be distributed for system development. Segments from the development and test set will be formatted with 8-bit (a-law) SPHERE files sampled at 8kHz.

The test set will be distributed by NIST via the online evaluation platform (https://lre.nist.gov).

4.1 Data Organization

The development and test sets follow a similar directory structure:

```
<base_directory>/
    README.txt
    data/
        dev/
        eval/
    docs/
    metadata/ (in dev set only)
```

4.2 Trial File

The trial file named `lre22_{dev|eval}_trials.tsv` and located in the `docs/` directory is composed of a header and a list of test segments:

```
segmentid<NEWLINE>
<segmentid><NEWLINE>
...
```

For example:

```
segmentid
1001_lre22
1002_lre22
1003_lre22
```

5 Evaluation Rules and Requirements

LRE22 is conducted as an open evaluation where the test data is sent to the participants who process the data locally and submit the output of their systems to NIST for scoring. As such, the participants must agree to process the data in accordance with the following rules:

- The participants agree that for each evaluation test segment the information available to the system is limited to that segment only (along with the training data); scores for a particular test segment must be computed without benefit from any information that might be derived from other test segments.
- The participants agree not to probe the test segments via manual/human means such as listening to the data or producing the transcript of the speech during the evaluation period and before all submissions are made.
- The participants are allowed to use information available in the audio file header.

In addition to the above data processing rules, participants agree to comply with the following general requirements:

- The participants agree to follow the submission requirements. See Section 6.4.
- The participants agree to have one or more representatives at the evaluation workshop to present a meaningful description of their system(s). Evaluation participants failing to do so will be excluded from future evaluation participation.
- The participants agree to the guidelines governing the publication of the results:
  - Participants are free to publish results for their own system but must not publicly compare their results with other participants (ranking, score differences, etc.) without explicit written consent from the other participants.
  - While participants may report their own results, participants may not make advertising claims about winning the evaluation or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (14 C.F.R. § 200.113) shall be respected[5]: *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*
  - At the conclusion of the evaluation NIST generates a report summarizing the system results for conditions of interest, but these results/charts do not contain the participant names of the systems involved. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source.
  - The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

6 Evaluation Protocol

To facilitate efficient information exchange between the participants and NIST, all evaluation activities are conducted over a web-interface.

6.1 Evaluation Account

Participants must sign up for an evaluation account where they can perform various activities such as registering for the evaluation, signing the data license agreement, uploading the submission and system description, and more. To sign up for an evaluation account, go to https://lre.nist.gov. The password must be at least 12 characters long and must contain a mix of upper and lowercase letters, numbers, and symbols. After the evaluation account is confirmed, the participant is asked to join a site or create one if it does not exist. The participant is also asked to associate his or her site to a team or create one if it does not exist. This allows multiple members with their individual accounts to perform activities on behalf of their site and/or team (e.g., making a submission) in addition to performing their own activities (e.g., requesting workshop invitation letter). Please note that the first person that creates the site or team is deemed the team representative. Site and team representatives have to approve participants who want to join his/her site/team:

- A site is defined as a single organization (e.g., NIST).

[5]See http://www.ecfr.gov/cgi-bin/ECFR?page=browse

- A team is defined as a group of sites (or organizations) collaborating on a task (e.g., Team1 consisting of NIST and LDC).
- A participant is defined as a member or representative of a site who takes part in the evaluation (e.g., John Doe).

6.2 Evaluation Registration

One representative from the team[6] must formally register his team to participate in the evaluation by agreeing to the terms of participation. For more information about the terms of participation, see Section 5.

6.3 Data License Agreement

One representative from each site must sign the LDC data license agreement to obtain the training data for the *fixed* training condition and Babel data for the *open* training condition.

6.4 Submission Requirements

Each team must participate in the *fixed* training condition. Teams are encouraged to participate in the *open* training condition to demonstrate the gains that can be achieved leveraging unconstrained amounts of data. There is no submission limit, but for each training condition participating teams must designate one submission as the *primary* submission that NIST can use for cross-team comparisons.

There should be one output file per training condition per system. Teams must process all test segments. Submissions with missing test segments will not pass validation and will be rejected.

Each team is required to submit a system description at the designated time (see Section 7). The evaluation results are made available only after the system description report is received and confirmed to comply with guidelines described in Section 6.4.1.

The system output file is composed of a header and a set of records where each record contains a test segment given in the file list (see Section 4.2) and a 14-dimension vector of log-likelihood scores. The order of the test segments in the system output file must follow the same order as the file list. Each record is a single line containing 14+1 fields, separated by tab character, in the order listed below:

TBD Segment ID<TAB>

1. afr-afr<TAB>
2. ara-aeb<TAB>
3. ara-arq<TAB>
4. ara-ayl<TAB>
5. eng-ens<TAB>
6. eng-iaf<TAB>
7. fra-ntf<TAB>
8. nbl-nbl<TAB>
9. orm-orm<TAB>
10. tir-tir<TAB>

[6]Please note that the registration is done at the team level while the data license is done at the site level. If a team is registered, all sites in that team are registered. However, all sites in the team must sign the data license separately.

11. tso-tso<TAB>
12. ven-ven<TAB>
13. xho-xho<TAB>
14. zul-zul<NEWLINE>

For example:

```
segmentid afr-afr ara-aeb ara-arq ...  zul-zul[7]
1001_lre22 -0.10017 -0.61518 -1.98380 ...  -2.47851
1002_lre22 -0.15862 -0.35402 -0.04077 ...  -0.96342
1003_lre22 -0.53162 -0.46526 -0.98556 ...  -1.23140
```

There should be one output file for each training condition for each system. System outputs will be automatically validated through the online submission platform and a report will be generated and displayed in case there are any errors.

6.4.1 System Description Format

Each team is required to submit a system description. The system description must include the following items:

- a complete description of the system components, including front-end (e.g., speech activity detection, features, normalization) and back-end (e.g., background models, embedding extractor, classifier) modules along with their configurations (i.e., filterbank configuration, dimensionality and type of the acoustic feature parameters, as well as the acoustic model and the backend model configurations),
- a complete description of the data partitions used to train the various models (as mentioned above). Teams are encouraged to report whether and how having access to the development set helped improve the performance,
- performance of the submission systems (primary and secondary) on the LRE22 development set, using the scoring software provided via the web platform (https://lre.nist.gov). Teams are encouraged to quantify the contribution of their major system components that they believe resulted in significant performance gains,
- a report of the CPU or GPU execution time (single threaded) and the amount of memory used to process a single trial (i.e., the time needed for processing a test segment to compute the score vector).

The system description should follow the latest IEEE ICASSP conference proceeding template.

7 Schedule (Tentative)

Milestone	Date
Evaluation plan published	August 2022
Registration period	September - October 2022
Training & development data available	September, 2022
Test data available to participants	October 17, 2022
System output due to NIST	November 18, 2022
Preliminary results released	December 9, 2022
Post evaluation workshop	January 31, 2023

[7]Note that the header is in lower case and output files without the header will not pass the validation step.

8 Acknowledgement

NIST has received support from other U.S. government agencies, such as Department of Defense, Department of Justice, and Intelligence Advanced Research Projects Activity (IARPA), to build a forum for the advancement of language recognition technology through the NIST LRE series.

References

[1] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, J. Hernandez-Cordero *et al.*, "The 2017 nist language recognition evaluation." in *Odyssey*, 2018, pp. 82–89.

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

[3] L. J. Rodriguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, G. Bordel, and M. Diez, "The albayzin 2012 language recognition evaluation." in *Interspeech*, 08 2013, pp. 1497–1501.

[4] J. Fiscus, W. M. Fisher, A. F. Martin, M. A. Przybocki, and D. S. Pallett, "2000 nist evaluation of conversational speech recognition over the telephone: English and mandarin performance results," in *Proc. Speech Transcription Workshop*. Citeseer, 2000, pp. 1–5.