Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

The 2017 NIST Language Recognition Evaluation

Seyed Omid Sadjadi[1,], Timothée Kheyrkhah[1,†], Audrey Tong[1], Craig Greenberg[1], Douglas Reynolds[2], Elliot Singer[2], Lisa Mason[3], Jaime Hernandez-Cordero[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

craig.greenberg@nist.gov

Abstract

In 2017, the U.S. National Institute of Standards and Technology (NIST) conducted the most recent in an ongoing series of Language Recognition Evaluations (LRE) meant to foster research in robust text- and speaker-independent language recognition as well as measure performance of current state-of-the-art systems. LRE17 was organized in a similar manner to LRE15, focusing on differentiating closely related languages (14 in total) drawn from 5 language clusters, namely Arabic, Chinese, English, Iberian, and Slavic. Similar to LRE15, LRE17 offered *fixed* and *open* training conditions to facilitate cross-system comparisons, and to understand the impact of additional and unconstrained amounts of training data on system performance, respectively. There were, however, several differences between LRE17 and LRE15 most notably including: 1) use of audio extracted from online videos (AfV) as development and test material, 2) release of a small development set which broadly matched the LRE17 test set, 3) system outputs in form of log-likelihood scores, rather than log-likelihood ratios, and 4) an alternative cross-entropy based performance metric. A total of 25 research organizations, forming 18 teams, participated in this 1-month long evaluation and, combined, submitted 79 valid system outputs to be evaluated. This paper presents an overview of the evaluation and an analysis of system performance over all primary evaluation conditions. The evaluation results suggest that 1) language recognition on AfV data was, in general, more challenging than telephony data, 2) top performing systems exhibited similar performance, 3) greatest performance improvements were largely due to data augmentation and use of more complex models for data representation, and 4) effective use of the development set was essential for the top performing systems.

1. Introduction

The National Institute of Standards and Technology (NIST) organized the 2017 Language Recognition Evaluation (LRE17) in the fall of 2017. The LRE17 was the latest in the ongoing series of language recognition technology evaluations conducted by NIST since 1996 [1]. The objectives of the evaluation series are 1) to stimulate and explore promising new ideas in robust text- and speaker-independent language recognition, 2) to support the development of advanced technology incorporating these ideas, and 3) to measure and calibrate the performance of the current state of technology. Figure 1 shows the number of target languages and participants in all NIST LREs organized to date. Although there seems to be a decrease in the number of participants for LRE17 compared to LRE15, we saw more collaborations this year where several previous participants formed new teams or joined existing teams.

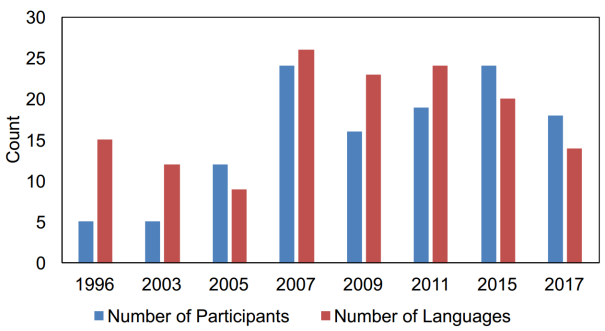

Figure 1: *Target language and participant statistics of the NIST LRE series.*

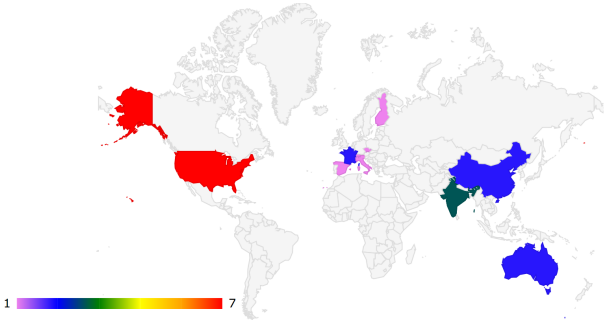

Figure 2: *Heat map of the world countries showing the number of LRE17 participating sites per country.*

The basic task in the NIST LREs is language detection, that is, determining whether a specified target language is spoken in a given test speech recording. Since LRE11 [2], the focus of the language detection task has shifted towards differentiating closely related languages that are sometimes mutually intelligible.

LRE17 was organized entirely online in a similar manner to LRE15 [3], using a web platform deployed on Amazon Web Services (AWS)[1]. The web platform supported a variety of services including evaluation registration, software and

*Contractor, †Guest Researcher

[1]see Disclaimer.

10.21437/Odyssey.2018-12

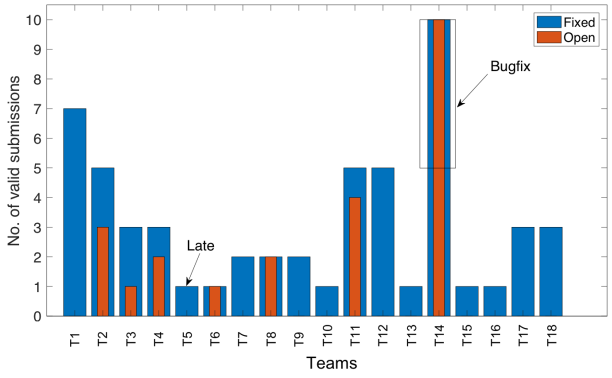

Figure 3: *LRE17 submission statistics.*

data distribution, system output submission, submission validation/scoring, and system description/presentation uploads. The online platform made LRE17 more readily accessible, and a total of 25 sites, forming 18 teams, from 11 countries registered for the evaluation. Figure 2 displays a heatmap representing the number of sites per country. It should be noted that all participant information, including country, was self-reported.

Similar to LRE15, LRE17 offered *fixed* and *open* training conditions. In the *fixed* training condition, participants were only allowed to use pre-specified "common" training data to develop their systems and build target language models. On the other hand, in the *open* training scenario, additional publicly available or proprietary data was permitted for use in system and model development. The inclusion of non-publicly available or proprietary data was new for LRE17. System output submission for the *fixed* training condition was required for all LRE17 participants to allow meaningful cross-system comparisons, while submission to the *open* training condition was optional but strongly encouraged to help quantify the contribution of unconstrained amounts of data on system performance. The number of submissions per team in LRE17 is shown in Figure 3. A total of 79 valid submissions were received, 56 of which were for the *fixed* training condition, and the remaining 23 were for the *open* training condition.

There were, however, several differences between LRE17 and LRE15. First, in addition to conversational telephone speech (CTS) and broadcast narrow band speech (BNBS), audio extracted from online videos (AfV) were used as development and test material in LRE17. The test segments from CTS and BNBS sources, extracted from longer recordings, were chunked to contain approximately 3 s, 10 s, or 30 s of speech as determined via an automatic speech activity detector. The test segments from AfV used the entire recording, and had durations ranging from 10 s to 900 s. Second, a small, yet representative, development set was released along with the training set. This *Dev* set, which broadly matched the LRE17 test set, could be used for both system training and development (e.g., hyperparameter tuning) purposes. Third, unlike in LRE15, systems were required to provide a vector of log-likelihood scores, rather than log-likelihood ratios, which provided the opportunity for a more in-depth system performance analysis (e.g., cross-year performance comparison). Fourth, the primary performance metric in LRE17, which was an average of costs calculated at two operating points, supported equal weighting of data sources and segment durations. In addition, an alternative performance metric, termed normalized cross-entropy (NCE) [4], was adopted in LRE17. Finally, NIST released to LRE17 participants an i-vector based language recognition system to serve as a reproducible state-of-the-art (as of LRE15) baseline, as well as to lower the barrier to entry for those participants having access to limited resources for building state-of-the-art systems.

2. Data

In LRE17, performance was evaluated by presenting systems with a series of test and target-language speech recordings. There were a total of 14 target languages drawn from 5 language clusters, namely Arabic, Chinese, English, Iberian, and Slavic. Unlike in LRE15 that focused on distinguishing languages within each cluster, LRE17 used both intra- and inter-cluster languages as non-targets. Table 1 shows the target languages (along with the language codes [5]) and corresponding language clusters in LRE17.

Cluster	**Target Language (code)**
Arabic	Egyptian Arabic (ara-arz), Iraqi Arabic (ara-acm), Levantine Arabic (ara-apc), Maghrebi Arabic (ara-ary)
Chinese	Mandarin (zho-cmn), Min Nan (zho-nan)
English	British English (eng-gbr), General American English (eng-usg)
Slavic	Polish (qsl-pol), Russian (qsl-rus)
Iberian	Caribbean Spanish (spa-car), European Spanish (spa-eur), Latin American Continental Spanish (spa-lac), Brazilian Portuguese (por-brz)

Table 1: *LRE17 target languages and language clusters.*

In this section we provide a brief description of the data used in LRE17 for training, development, and test.

2.1. Training set

As noted previously, there were two training conditions in LRE17, namely *fixed* and *open*. The *fixed* training condition limits the system training to the following specific data sets, which were made available to the participants by the Linguistic Data Consortium (LDC): i) previous LRE data (as released in LDC2017E22), ii) Fisher English corpus (LDC2004S13 [6], LDC2004T19 [7], LDC2005S13 [8], LDC2005T19 [9]), iii) Switchboard (SWB) corpora (LDC97S62 [10], LDC98S75 [11], LDC99S79 [12], LDC2001S13 [13], LDC2002S06 [14], LDC2004S07 [15]), and iv) LRE17 *Dev* set (LDC2017E23).

Figure 4 shows the number of speech segments available in LDC2017E22 for each target language. Among all the languages, Levantine Arabic (ara-apc) had the most recordings (3509), while Chinese Min Nan had the least (95). On average, there were 1157 speech segments per target language. From a total of 16,205 segments in LDC2017E22, 13,956 were selected from CTS recordings, and the remaining 2249 were from BNBS recordings. Figure 5 depicts the source type distribution of the target language training data. Training data for most languages (all except for British English, General American English, Brazilian Portuguese, Polish, and Russian, and Chinese Mandarin) were drawn from a single source type, which was predominantly CTS.

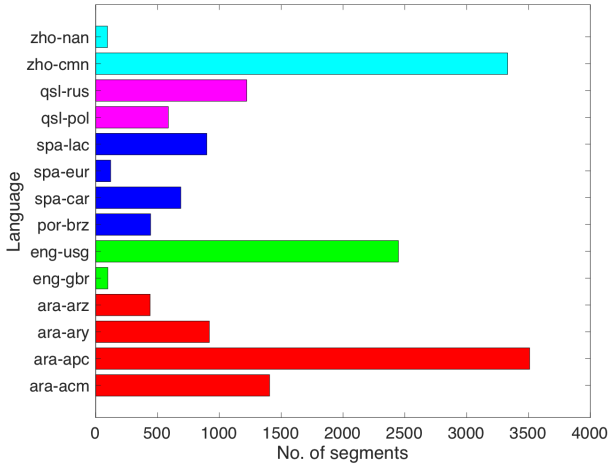

Figure 4: *LRE17 training set counts per target language.*

Switchboard and Fisher corpora were included here because they contain transcripts, making them suitable for training acoustic models, e.g., deep neural network (DNN) models. In addition to these, publicly available non-speech audio and data (e.g., noise and non-vocal music samples, impulse responses, filters) could be used for system training and development purposes. Participation in the *fixed* condition was required. It is worth noting that the use of pretrained models on data other than what was designated above was also not allowed in the *fixed* condition.

In the *open* training scenario, additional data, including proprietary data and data that are not publicly available, was permitted for use in system training and development. The inclusion of non-publicly available data was new in LRE17. LDC also made available selected data from the IARPA Babel Program [16] to be used in the *open* training condition. Participation in this condition was optional but strongly encouraged to help quantify the gains that one could achieve with unconstrained amounts of data.

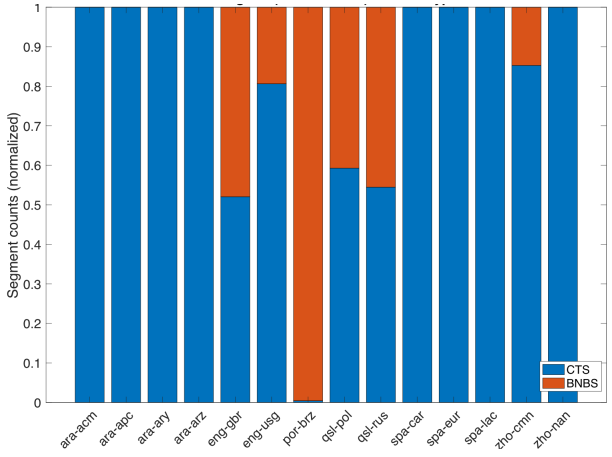

Figure 5: *Distribution of source type (CTS vs BNBS) for training languages in LRE17 (LDC2017E22).*

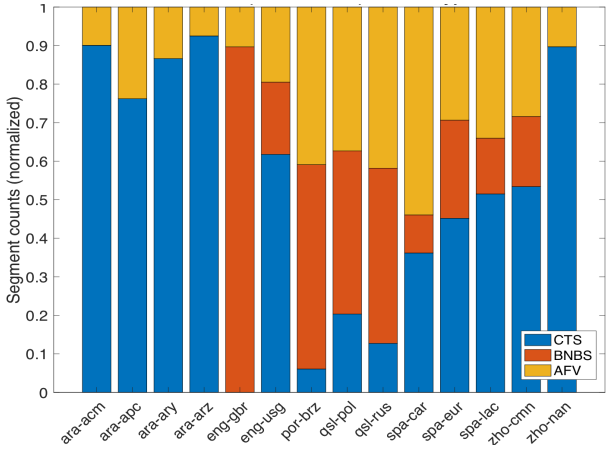

Figure 6: *Number of segments per target language in LRE17 Dev set (LDC2017E23) by source type.*

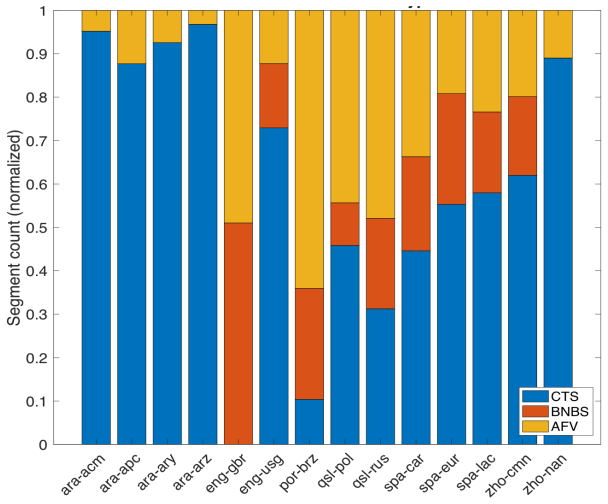

Figure 7: *Number of segments per target language in LRE17 test set by source type.*

2.2. Development and test sets

The speech segments in the LRE17 *Dev* and *test* sets were extracted from Multi-language Speech (MLS14) [17] and Video Annotation for Speech Technologies (VAST) [18] corpora, both of which were collected by the LDC to support speech technology evaluations. MLS14 consists of CTS and BNBS recordings, while VAST contains AfV data. Original speech sessions from MLS14 were split into nested (i.e., overlapping) 3 s, 10 s, and 30 s cuts based on speech activity detection (SAD) marks. Only one nested 30 s/10 s/3 s cut was selected per recording session, and there were equal number of cuts in each duration bin. Segments from the MLS14 corpus were encoded as μ-law (8-bit) sampled at 8 kHz in NIST SPHERE [19] formatted files, while recordings from the VAST corpus were 16-bit FLAC files sampled at 44 kHz.

Figures 6 and 7 show the source type distributions of the LRE17 *Dev* and *test* segments by language, respectively. It can be seen from the figures that a majority of the segments were drawn from one source type, which, like the training data, was predominantly CTS (except for British English that is domi-

nated by BNBS). Also, both sets contain speech segments from the AfV source type for all target languages. From a total of 3661 segments in LRE17 *Dev* set (LDC2017E23), 1999 were selected from CTS recordings, 788 were from BNBS recordings, and the remaining 874 from AfV. As for the LRE17 *test* set, from a total of 25,451 cuts, 15,018 were extracted from CTS recordings, 2002 were from BNBS recordings, and the remaining 3521 from AfV.

3. Performance measurements

As noted in Introduction, systems submitted to LRE17 were required to provide a 14-dimensional vector of log-likelihood scores, corresponding to the 14 target languages, for each test segment. 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.

3.1. Primary metric

In LRE17, pair-wise language recognition performance was computed for all target-language/non-target-language pairs (L_T, L_N). This was done in terms of false-reject (missed detection) and false alarm (FA) probabilities, which were computed separately for each target language and each target/non-target language pair, respectively. The miss and false alarm probabilities were 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 spurious detection), and P_{Target} (*a priori* probability of the specified target language) are the 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: *LRE17 cost parameters.*

Note that the first set of parameter values are those historically used in the past NIST LREs and provide equal weighting to miss and false alarm errors, while the second set of parameters are not similarly balanced. Therefore, to improve the interpretability of the cost function, it was 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

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

The default cost for both sets of parameters defined in Table 2 was 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 were computed by applying detection thresholds of $\log(\beta)$ to log-likelihood *ratios* derived from the log-likelihoods output by the system[2].

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

$$\begin{aligned} C_{avg}(\beta) = \frac{1}{N_L}\Bigg\{ & \sum_{L_T} P_{Miss}(L_T) + \\ & \frac{1}{N_L - 1}\Big[\beta \times \sum_{L_T}\sum_{L_N} P_{FA}(L_T, L_N)\Big]\Bigg\}, \end{aligned} \tag{6}$$

where N_L is the number of target languages. The primary metric for LRE17 was the average cost performance defined in (6), computed using the two application model parameters given in Table 2, that were then averaged:

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

Unlike in previous LREs, in LRE17 the evaluation data was divided into partitions based on the data source, i.e., MLS14 and VAST, for each language, resulting in a total 28 partitions (2×14). In other words, for each language, the counts for each corpus (MLS14 and VAST) were equalized. C_{avg} was calculated for each partition, and the final result was the average of all the partitions' C_{avg}'s. The average of basic C_{avg} scores for the two set of parameters defined in Table 2 served as the primary metric to measure a system performance. Also, the minimum detection cost, *min*$C_{primary}$, was computed by using the detection thresholds that minimize the detection cost. Note that for minimum cost calculations, the counts for each condition set were equalized before pooling and cost calculation (i.e., minimum cost was computed using a single threshold not one per condition set).

NIST released to LRE17 participants a software package that supported validation of the system outputs (to ensure they conformed to formatting guidelines provided in the evaluation plan) as well as calculation of the primary metric.

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

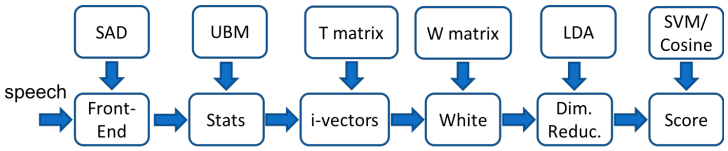

Figure 8: *Block diagram of the NIST LRE17 baseline system.*

3.2. Alternative metric

In addition to the cost metric $C_{primary}$ described in Section 3.1, an alternative information theoretic performance measure was also used to calculate the performance of language recognition systems in LRE17. The alternative measure, termed Normalized Cross-Entropy (aka confidence score) [4], has been used by NIST since late 90's to evaluate performance of speech-to-text systems. In this section, we provide a brief description of NCE as well as its interpretation for language recognition tasks.

The multiclass cross-entropy H_{mce} measures the information a language recognition system provides through the log-likelihood scores and is defined as follows [20]

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

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 language recognition 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

$$\text{NCE} = \frac{H_{max} - H_{mce}}{H_{max}}. \tag{10}$$

Given that the cross-entropy is non-negative, a perfect language recognition 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). The idea of this *normalized* cross-entropy measure is to allow for comparison of confidence scores across different test sets by correcting for the effect of prior probabilities, which tend to boost raw cross-entropy scores when they are high.

It is worth noting that the confidence metric in (10) is being considered for use as the primary metric in future LREs.

4. Baseline system

As noted in Introduction, NIST released a baseline language recognition system to LRE17 participants. The system was developed using the NIST SLRE toolkit, and meant to 1) serve as a baseline for the current state of technology in language recognition which is reproducible by all participants, and 2) lower the barrier to entry for those participants who may have access to limited resources for building state-of-the-art systems.

A schematic block diagram of the system is depicted in Figure 8. It supports both frontend processing (e.g., SAD, feature extraction and normalization) and backend modeling, e.g., Gaussian mixture model (GMM) training, i-vector extraction, linear discriminant analysis (LDA), for language recognition. It also provides tools for the extraction of DNN Bottleneck features (BNF). For the BNFs, NIST also made available pre-trained DNN models built using Kaldi [21] on a senone set with nearly 8700 targets obtained from *tri5a* stage in Kaldi's *fisher-swbd* example. The models were trained on speech data from combined SWB1 and Fisher corpora (∼2000 hours), using hidden layers (hidden units: 2048-2048-2048-2048-2048-80-1024) with REctified Linear Unit (RELU) activation followed by a renorm nonlinearity that scales the RMS of the vector of activations to 1.0. The bottleneck layer (second to the output), which has 80 hidden units, only uses the renorm nonlinearity. A 21-frame context of 39-dimensional (13 static + Δ + Δ^2) mel-frequency cepstral coefficients (MFCCs), extracted using the NIST SLRE toolkit, was used as input to the DNN. The MFCCs were extracted from 25 ms frames every 10 ms using a 24-channel mel filterbank spanning the frequency range 100 Hz-4000 Hz.

System	BNF	UBM	T	Wht/LDA	Cos/SVM
Dev	Fisher + SWB1	E22	E22	E22	E22
Eval	Fisher + SWB1	E22	E22	E22+E23	E22+E23

Table 3: *Summary of datasets used to train the various components of the NIST LRE17 baseline system.*

For non-speech frame dropping, a statistical model based SAD [22] was adopted. After dropping the non-speech frames, segment level cepstral mean and variance normalization (CMVN) was applied and followed by short-time cepstral mean subtraction over a 3-second sliding window.

For i-vector extraction, the system used a 500-dimensional total variability subspace (denoted as **T** in Table 3) trained on all speech segments from LDC2017E22. A 2048-component GMM-UBM with diagonal covariance matrices, also trained on LDC2017E22, was used to compute the zeroth and first order Baum-Welch statistics. Before dimensionality reduction through LDA, the 500-dimensional i-vectors were whitened using within-class covariance normalization (WCCN), and unit-length normalized. For backend scoring, the system supported two commonly adopted techniques, i.e., cosine similarity measure, and support vector machines (SVM) with a Gaussian kernel.

Two different configurations were used for the baseline system, namely Dev and Eval. In the Dev configuration, the LRE17 *Dev* set (i.e., LDC2017E23) was excluded from system training,

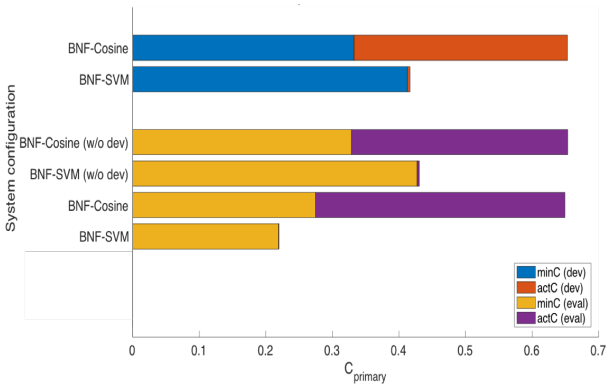

Figure 9: *NIST LRE17 baseline system performance.*

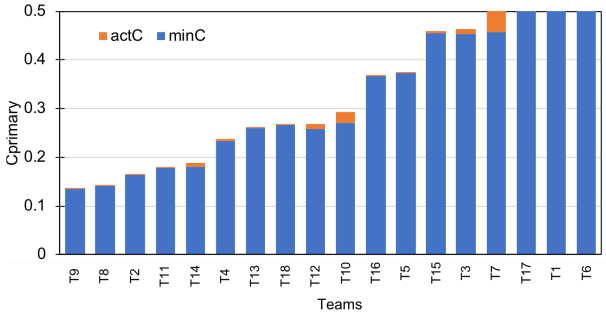

Figure 10: *Actual and minimum costs for LRE17 primary **fixed** submissions.*

and it was used as the test set, while in the Eval configuration, the *Dev* set was included for training the whitening and LDA transforms as well as the classifiers. A summary of the datasets used to train the various components of the NIST LRE17 baseline system is given in Table 3.

It is worth emphasizing that the configuration parameters employed to build the LRE17 baseline system are commonly used by the language recognition community, and no attempt was made to tune the hyperparameters or data lists utilized to train the models.

5. Results

In this section, we present results and performance analysis of the baseline system as well as all LRE17 primary submissions, in terms of minimum and actual $C_{Primary}$, and NCE.

Figure 9 shows the actual and minimum costs for the NIST LRE17 baseline system obtained using Dev (blue/red bars) and Eval (yellow/purple bars) configurations discussed in Section 4. Two important observations can be made from this figure. First, without using the LRE17 *Dev* set (i.e., LDC2017E23) in system training, similar performances are observed on the *Dev* and *test* sets. Second, including the LRE17 *Dev* set in training remarkably improves the performance, in particular for the SVM classifier.

Figure 10 shows the actual and minimum costs for all primary submissions in the *fixed* training condition. Here, the y-axis upper limit is set to 0.5 to facilitate cross-system comparisons in the lower $C_{Primary}$ region. It can be seen from the figure that the performance gap among the top-5 teams is not

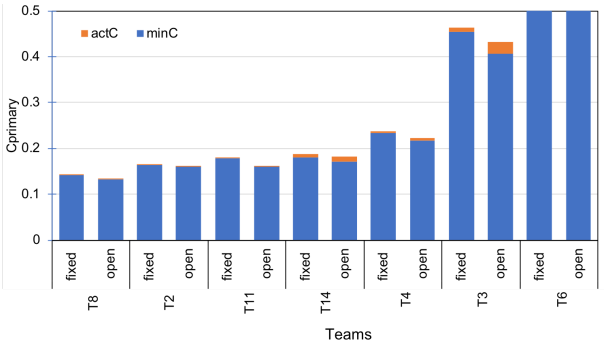

Figure 11: *Impact of **fixed** (left) versus **open** (right) training on language recognition performance in terms of actual and minimum cost.*

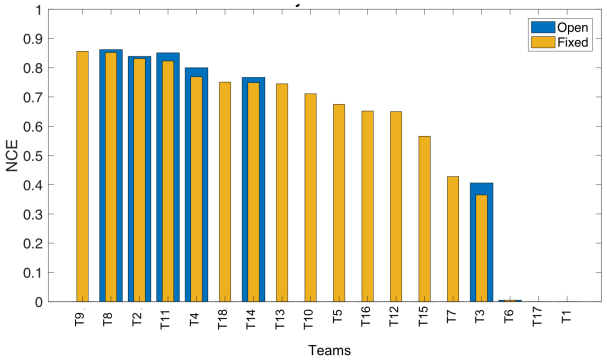

Figure 12: *Performance in terms of NCE for LRE17 primary **fixed** and **open** submissions.*

remarkable. It is also observed that score calibration was successfully applied for almost all teams (i.e., the absolute difference between the minimum and actual costs is relatively small).

Figure 11 shows system performance by training condition for the 7 teams that participated in both *fixed* and *open* tasks. We observe limited improvement in the *open* training condition over the *fixed* training condition. However, unlike in LRE15 where in some cases worse performance was observed for the *open* training conditions, we see consistent, though small, improvement with additional data for all primary *open* submissions.

Figure 12 shows the NCE performance measure (i.e., LRE17 alternative metric) for all primary submissions in both *fixed* and *open* training conditions (higher score is better). Here, the y-axis lower limit is set to 0 for better visualization of the results in the higher NCE region. Overall, we observe a similar performance trend as with the primary metric, in particular for the top performing systems. There are, however, some changes in the position of teams on the bar plot.

Figure 13 shows the results in terms of actual $C_{Primary}$ for the top four performing primary submissions under the *fixed* training condition. We observe that costs vary widely for different target languages. For example, the detection cost for `spa-car` is more than 10 times worse (larger) than the cost for `qsl-rus`. More generally, performance on the Slavic target languages (i.e., `qsl-pol` and `qsl-rus`) tends to be the best, while performance on the Iberian target languages (in particular, `spa-car`, `spa-lac`) tends to be the worst. Another

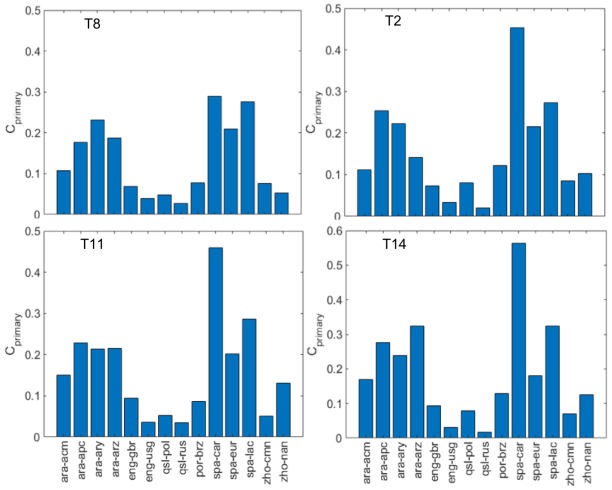

Figure 13: *Performance by language in terms of actual cost for the top four primary **fixed** submissions.*

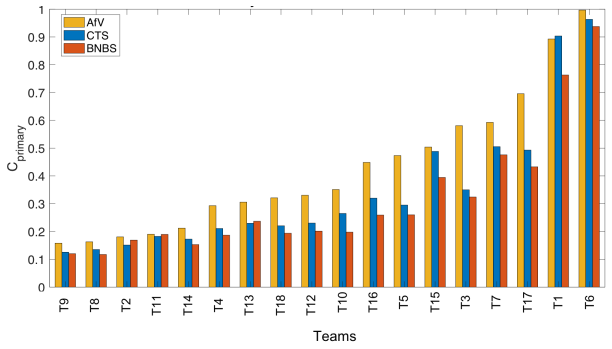

Figure 14: *Performance by source type in terms of actual cost for LRE17 primary **fixed** submissions.*

interesting observation from the figure is that the performance trend for the top performing system (upper left panel) on some target languages (e.g., `spa-car`, `zho-nan`, and `ara-apc`) is different than the performance trend seen with the other three systems.

Figure 14 shows the results in terms of actual $C_{Primary}$ based on test segment source type (i.e., CTS vs BNBS vs AfV), where AfV represents itself as the most challenging source type in LRE17. This is expected because 1) diverse sources of interference (background noise, music, competing speakers, etc) are typically present in online videos, and 2) there is a dramatic domain mismatch between telephony data, which is very well represented in the training set, and AfV, which is sparse. Furthermore, small performance difference is observed between CTS and BNBS segments. For telephony data, system performance on CTS segments is somewhat worse than that on BNBS, and we speculate this is due to which languages had predominantly CTS test segments (for instance, target languages drawn from Arabic and Iberian language clusters).

In Figure 15, we see performance for all primary *fixed* submissions broken down by speech duration. There are equal number of segments in each duration bin. Here, we only report the results on the MLS14 portion of the test set because for the VAST portion enitre recordings with varying speech dura-

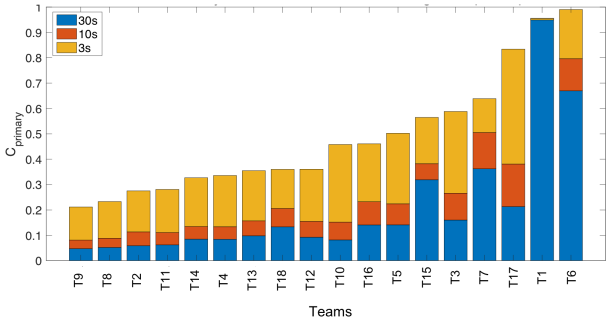

Figure 15: *Performance by speech duration (MLS14) in terms of actual cost for primary **fixed** submissions.*

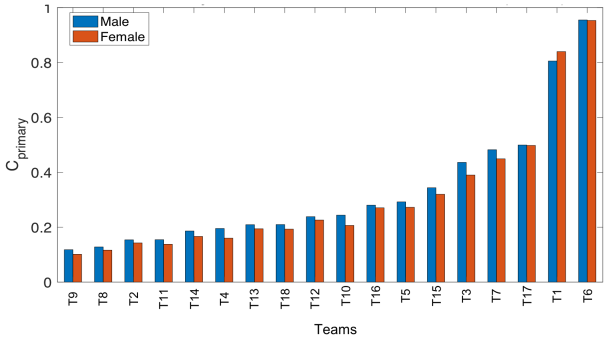

Figure 16: *Performance by speaker gender (MLS14) in terms of actual cost for primary **fixed** submissions.*

tions were used. It can be seen from the figure that performance rapidly improves as speech duration increases from 3 seconds to 10 seconds. The improvement is more substantial for an increase in speech duration from 3 s to 10 s than 10 s to 30 s.

Figure 16 shows the results in terms of actual cost based on speaker gender of test segments. Here, we only report the results on the MLS14 portion of test set that included gender metadata. Little performance difference is observed between male and female speakers, and segments from male speakers seem to be more challenging for language recognition than segments from female speakers. It is worth noting that these results are in line with LRE15 observations.

6. Conclusion

This paper presented a summary of the 2017 NIST language recognition evaluation, whose objective was to provide a platform for evaluating the most advanced technology in language recognition and to foster new ideas and collaboration. LRE17 attracted worldwide research organizations from academia and industry, including 7 first time participants.

LRE17 introduced several new aspects, most importantly: 1) release of a small development set which broadly matched the LRE17 test set, 2) use of audio extracted from online videos as development and test material, 3) system outputs in form of log-likelihood scores, rather than log-likelihood ratios, and 4) an alternative cross-entropy based performance metric. NIST also released a language recognition system to serve as a baseline for the current state of technology, and to lower the barrier to entry for the evaluation.

It was observed that, overall, language recognition on AfV

data was more challenging than on telephony speech. Additionally, we saw that for some target languages (e.g., `spa-car`), the top performing system has significantly better performance than the rest of the systems. Another insight from this evaluation is that unconstrained amounts of data (either publicly available or proprietary) under *open* training condition does not seem to lead to substantially better performance. Our plan is to report on additional analysis of system performance results in the near future.

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.

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.

The work of MIT Lincoln Laboratory is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

8. References

[1] NIST, "NIST Language Recognition Evaluation," https://www.nist.gov/itl/iad/mig/language-recognition, [Online; accessed 26-January-2018].

[2] NIST, "2011 Language Recognition Evaluation," https://www.nist.gov/itl/iad/mig/2011-language-recognition-evaluation, [Online; accessed 26-January-2018].

[3] NIST, "2015 Language Recognition Evaluation," https://www.nist.gov/itl/iad/mig/2015-language-recognition-evaluation, [Online; accessed 26-January-2018].

[4] NIST, "A tutorial introduction to the ideas behind Normalized Cross-Entropy and the information-theoretic idea of Entropy," https://www.nist.gov/file/411831, [Online; accessed 26-January-2018].

[5] SIL International, "Documentation for ISO 639 identifier," http://www-01.sil.org/iso639-3/, 2017, [Online; accessed 26-January-2018].

[6] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English Training Speech Part 1 Speech," https://catalog.ldc.upenn.edu/LDC2004S13, 2004, [Online; accessed 26-January-2018].

[7] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English Training Speech Part 1 Transcripts," https://catalog.ldc.upenn.edu/LDC2004T19, 2004, [Online; accessed 26-January-2018].

[8] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English Training Speech Part 2 Speech," https://catalog.ldc.upenn.edu/LDC2005S13, 2004, [Online; accessed 26-January-2018].

[9] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English Training Speech Part 2 Transcripts," https://catalog.ldc.upenn.edu/LDC2005T19, 2004, [Online; accessed 26-January-2018].

[10] J. Godfrey and E. Holliman, "Switchboard-1 Release 2," https://catalog.ldc.upenn.edu/LDC97S62, 1993, [Online; accessed 26-January-2018].

[11] D. Graff, A. Canavan, and G. Zipperlen, "Switchboard-2 Phase I," https://catalog.ldc.upenn.edu/LDC98S75, 1998, [Online; accessed 26-January-2018].

[12] D. Graff, K. Walker, and A. Canavan, "Switchboard-2 Phase II," https://catalog.ldc.upenn.edu/LDC99S79, 1999, [Online; accessed 26-January-2018].

[13] D. Graff, D. Miller, and K. Walker, "Switchboard-2 Phase III," https://catalog.ldc.upenn.edu/LDC2002S06, 2002, [Online; accessed 26-January-2018].

[14] D. Graff, K. Walker, and D. Miller, "Switchboard Cellular Part 1 Audio," https://catalog.ldc.upenn.edu/LDC2001S13, 2001, [Online; accessed 26-January-2018].

[15] D. Graff, K. Walker, and D. Miller, "Switchboard Cellular Part 2 Audio," https://catalog.ldc.upenn.edu/LDC2004S07, 2004, [Online; accessed 26-January-2018].

[16] M. P. Harper, "Data resources to support the Babel program," https://goo.gl/9aq958, [Online; accessed 26-January-2018].

[17] K. Jones, D. Graff, J. Wright, K. Walker, and S. Strassel, "Multi-language speech collection for NIST LRE," in *Proc. LREC*, Portoroz, Slovenia, May 2016, pp. 4253–4258.

[18] J. Tracey and S. Strassel, "VAST: A corpus of video annotation for speech technologies," in *Proc. LREC*, Miyazaki, Japan, May 2018.

[19] NIST, "Speech file manipulation software (SPHERE) package version 2.7," ftp://jaguar.ncsl.nist.gov/pub/sphere-2.7-20120312-1513.tar.bz2, 2012, [Online; accessed 26-January-2018].

[20] L. J. Rodríguez-Fuentes, N. Brümmer, M. Peñagarikano, A. Varona, G. Bordel, and M. Díez, "The Albayzin 2012 language recognition evaluation," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 1497–1501.

[21] D. Povey *et al.*, "Kaldi Speech Recognition Toolkit," https://github.com/kaldi-asr/kaldi, [Online; accessed 26-January-2018].

[22] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.