Performance Analysis of the 2017 NIST Language Recognition Evaluation

Seyed Omid Sadjadi[1,*], *Timothee Kheyrkhah*[1,†], *Craig Greenberg*[1],
Elliot Singer[2], *Douglas Reynolds*[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

The 2017 NIST Language Recognition Evaluation (LRE) was held in the autumn of 2017. Similar to past LREs, the basic task in LRE17 was language detection, with an emphasis on discriminating closely related languages (14 in total) selected from 5 language clusters. LRE17 featured several new aspects including: audio data extracted from online videos; a development set for system training and development use; log-likelihood system output submissions; a normalized cross-entropy performance measure as an alternative metric; and, the release of a baseline system developed using the NIST Speaker and Language Recognition Evaluation (SLRE) toolkit for participant use. A total of 18 teams from 25 academic and industrial organizations participated in the evaluation and submitted 79 valid systems under *fixed* and *open* training conditions first introduced in LRE15. In this paper, we report an in-depth analysis of system performance broken down by multiple factors such as data source and gender, as well as a cross-year performance comparison of leading systems from LRE15 and LRE17 to measure progress over the 2-year period. In addition, we present a comparison of primary versus "single best" submissions to understand the effect of fusion on overall performance.

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

1. Introduction

The 2017 NIST Language Recognition Evaluation (LRE17) was conducted in the autumn of 2017. It was the 8th cycle in the on-going language recognition technology evaluation series that began in 1996 [1], which serves to both stimulate and support research in robust language recognition as well as measure and calibrate the performance of language recognition systems. LRE17 was organized entirely online using a web platform[1] that supported a variety of evaluation related services such as registration, data license agreement submission, data distribution, system output submission, verification and scoring. The task in LRE17 was language detection, i.e., deciding whether or not a target language was spoken in a given test recording, with an emphasis on differentiating closely related languages drawn from 5 language clusters.

LRE17 had two training conditions, *fixed* and *open*, which were first introduced in LRE15. In the *fixed* scenario, NIST restricted system training and development data to specific data sets to facilitate meaningful cross-system comparisons in terms of core language recognition algorithms/approaches used. For the *open* condition, participants were allowed to explore the gains that could be obtained through the utilization of unconstrained amounts of publicly available and/or proprietary data. A total of 18 teams from 25 sites made 79 valid system submissions, 56 for the *fixed* training condition and 23 for the *open* training condition.

Compared to past LREs, LRE17 introduced several new features. First, in addition to conversational telephone speech (CTS) and broadcast narrow band speech (BNBS), audio extracted from YouTube[2] videos (AfV) was included in LRE17 as *test* data. Second, NIST provided participants with a small, yet representative, development (*dev*) set that broadly matched the LRE17 *test* set and could be used for both system training and development (e.g., hyperparameter tuning) purposes. Third, systems were required to provide a vector of log-likelihood scores, as opposed to the log-likelihood ratios required in LRE15, for each target language/test segment pair. The use of log-likelihoods gave NIST the opportunity for a more in-depth system performance analysis, such as the cross-year performance comparison presented later in this paper. Fourth, a new alternative performance metric, normalized cross-entropy (NCE) [2], was adopted in LRE17. The primary LRE17 performance metric used an average of costs calculated at two operating points, as in SRE16 [3], and supported equal weighting of data sources and segment durations [4]. However, the NCE metric not only measures the discrimination power of a language recognition system, it also reveals how well the system is calibrated. Finally, in an effort to lower the barrier to entry for LRE17 and provide a reproducible state-of-the-art system (as of LRE15), NIST released a baseline language recognition system developed using the NIST SLRE toolkit for participant use (for details regarding the baseline system and its performance see [5]).

Motivated by the observations from our initial analysis [5] of systems submitted for LRE17, in this paper we present a deeper analysis of system performance broken down by multiple factors such as data source and gender, as well as a cross-year (i.e., LRE15 versus LRE17) performance comparison of the top performing systems to measure progress in language recognition technology over the 2-year period. Additionally, given the emphasis on the development of "single best" systems in LRE17, we present a comparison of primary versus "single best" submissions to understand the impact of system fusion on overall performance.

2. Task description and target languages

The basic task in LRE17 was language detection, that is, given a segment of speech and a target language, automatically de-

[*] Contractor, [†] Guest Researcher

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

[2] See Disclaimer. YouTube is a trademark of Google LLC.

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.*

termine if the target language was spoken in the test segment. LRE17 included 14 target languages grouped into 5 language clusters, namely Arabic, Chinese, English, Iberian, and Slavic. Table 1 shows the target languages (along with the language codes [6]) and corresponding language clusters in LRE17.

Input to language recognition systems in LRE17 was a series of test segments, and the systems' output was a series of score vectors, one vector per test segment. Each 14-dimensional score vector contained estimated log-likelihood scores corresponding to the 14 target languages listed in Table 1. This is unlike LRE15 in which systems were required to submit log-likelihood ratios as scores.

3. Data

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

3.1. Training data

Similar to LRE15, LRE17 consisted of two training conditions: *fixed* and *open*. In the *fixed* training condition, system training and development data was limited to the following sets of data, which were made available to participants by the Linguistic Data Consortium (LDC):

- previous LRE data (as provided in LDC2017E22)
- Fisher English corpus [7, 8, 9, 10]
- Switchboard (SWB) corpora [11, 12, 13, 14, 15, 16]
- LRE17 *dev* set (LDC2017E23).

LDC2017E22 was the primary source of training data for the 14 target languages, containing a total of 16,205 segments: 13,956 from CTS recordings and 2,249 from BNBS recordings. It is worth noting here that the training data for most target languages were drawn from a single source type, which was predominantly CTS. Switchboard and Fisher corpora were included in the training set because they contain transcripts, making them suitable for training ASR acoustic models used by 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.

In the *open* training scenario, participants were allowed to utilize additional proprietary or publicly available data for system training and development. The inclusion of proprietary data was new in LRE17. Selected data from the IARPA Babel Program [17] was also made available by LDC to be used in the *open* training condition. Participation in this condition was optional but strongly encouraged to help quantify the gains that could be achieved with unconstrained amounts of data.

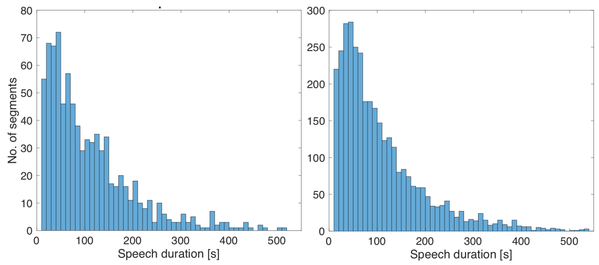

Figure 1: *Distribution of segment speech duration in the VAST portion of LRE17 dev (left) and test (right) sets.*

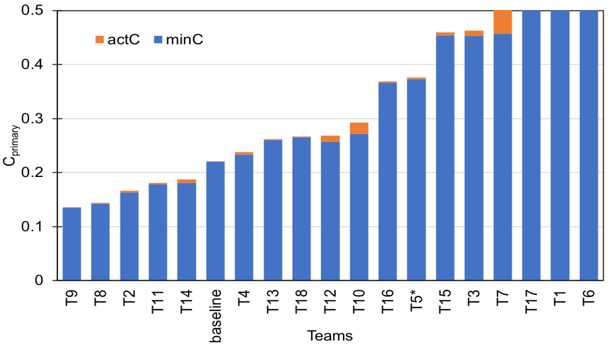

Figure 2: *Actual (actC) and minimum (minC) costs for LRE17 primary **fixed** submissions.*

3.2. Development and test sets

The speech segments in the LRE17 *dev* and *test* sets were extracted from Multi-language Speech (MLS14) [18] and Video Annotation for Speech Technologies (VAST) [19] 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 recordings from MLS14 were split into nested (i.e., overlapping) 3s, 10s, and 30s segments based on speech activity detection (SAD) marks. Only one nested 30s/10s/3s segment was selected per recording session, and there were equal number of segments in each duration bin. The test segments from AfV used the entire recording with speech durations ranging from 10s to 600s. Figure 1 shows distributions of segment speech durations in the LRE17 *dev* and *test* sets. It can be seen from the plots that the two sets have similar speech duration characteristics.

As in the training data, the majority of the *dev* and *test* segments were drawn from one source type, which was predominantly CTS. Also, both sets contained speech segments from the AfV source type for all target languages. From a total of 3,661 segments in LRE17 *dev* set (LDC2017E23), 1,999 were 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, 2,002 were from BNBS recordings, and the remaining 3,521 from AfV.

4. Results and analysis

In this section, we present performance analyses of LRE17 primary submissions, in terms of minimum and actual costs

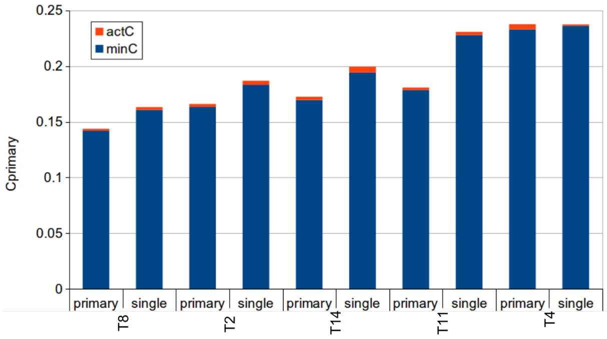

Figure 3: *Performance comparison of LRE17 primary versus "single best" submissions.*

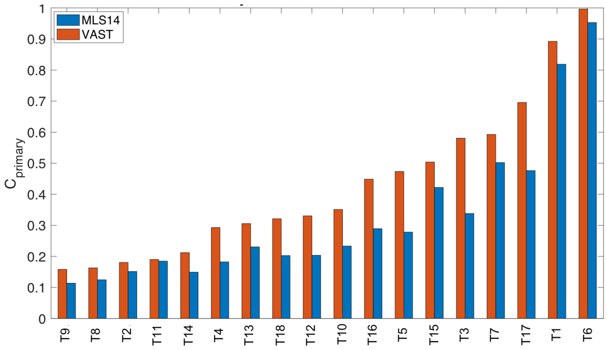

Figure 4: *Performance by data source in terms of actual cost for LRE17 primary* ***fixed*** *submissions.*

(aka $C_{Primary}$) that were the primary performance metrics in LRE17. We refer readers to the LRE17 evaluation plan [4] for details regarding the performance measures.

Figure 2 shows the actual and minimum costs for all primary submissions as well as the LRE17 baseline system (see [5] for details) in the *fixed* training condition. Here, the y-axis upper limit is set to 0.5 to facilitate cross-system comparisons in the lower cost region. Two important observations can be made from the figure; first, the performance gap among the top-5 teams is not substantial. Second, all teams seem to have successfully performed score calibration, i.e., the absolute difference between the minimum and actual costs is relatively small. It is worth noting that the T9 submission is simply a linear fusion (with equal weights) of T8 and T11 submissions, therefore it is excluded from our subsequent analyses.

There was a new special emphasis in LRE17 on the development of "single best" systems. Figure 3 shows the performance comparison of LRE17 primary versus "single best" systems in terms of minimum and actual costs for the participants who made "single" system submissions for LRE17. It can be seen from the figure that fusion still plays a notable role in the primary submissions, consistently providing gains for all systems. Nevertheless, it is observed that "single best" systems can perform as competitively as primary systems (e.g., the system submitted by T8).

Figure 4 shows system performance by data source (i.e., MLS14 versus VAST) in terms of actual cost for all LRE17 primary *fixed* submissions. As noted earlier, VAST corpus contains audio extracted from online videos (AfV), while MLS14 contains segments from CTS and BNBS source types. Overall, performance on VAST portion of the LRE17 test set seems to be worse than that on the MLS14 test segments. This is, however, not surprising given that the training and *dev* sets predominantly contain CTS/BNBS segments. Additionally, AfV data is expected to be more challenging because of the diverse sources of variability and distortion seen in YouTube videos.

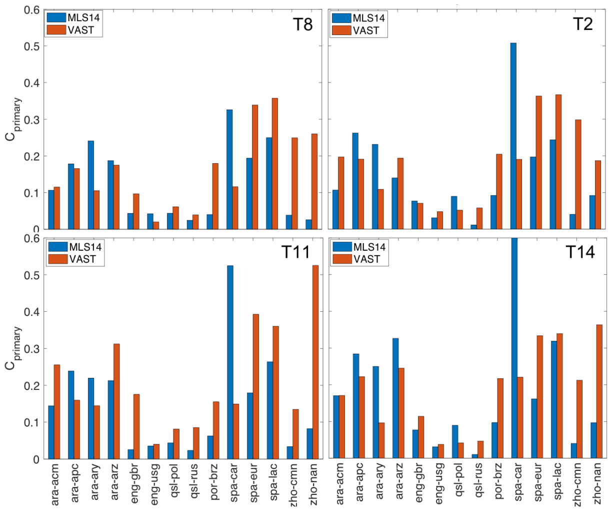

Figure 5: *Performance by data source per target language in terms of actual cost for the top four primary* ***fixed*** *submissions.*

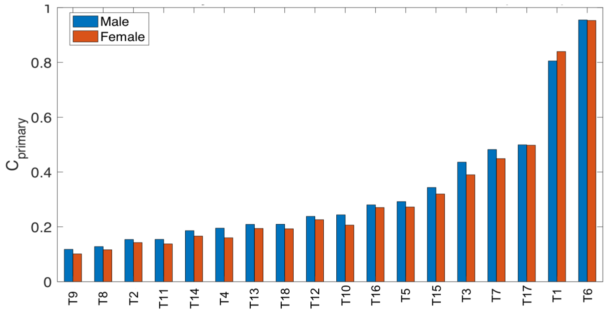

Figure 6: *Performance by speaker gender (MLS14) in terms of actual cost for primary* ***fixed*** *submissions.*

In order to gain further insight into the performance gap between the two data sources (i.e., MLS14 versus VAST), we analyzed the results presented in Figure 4 per target language for the top four primary *fixed* submissions. Figure 5 shows the outcome of this analysis. Interestingly, we see that for some target languages (in particular `spa-car`) performance is substantially worse on MLS14 test segments. It can also be seen that MLS14 versus VAST performance trends were not consistent for some target languages among the leading systems (e.g., Arabic cluster).

Figure 6 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 for which LDC provided gender metadata. A relatively small performance difference is observed between male and female speakers, with slightly worse performance for male speaker than female speakers. It is worth noting that these results are in line with LRE15 observations. We also analyzed performance by gender per target language, and the results are displayed in Figure 7. Performance on speaker gender varied with the target language, similar to our observations from Figure 5; particularly, the results on male segments were dramatically worse for `spa-car`.

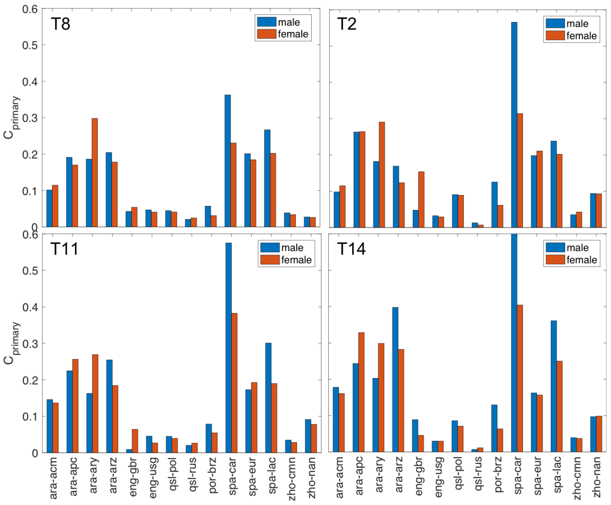

Figure 7: *Performance by gender per target language in terms of actual cost for the top four primary* ***fixed*** *submissions.*

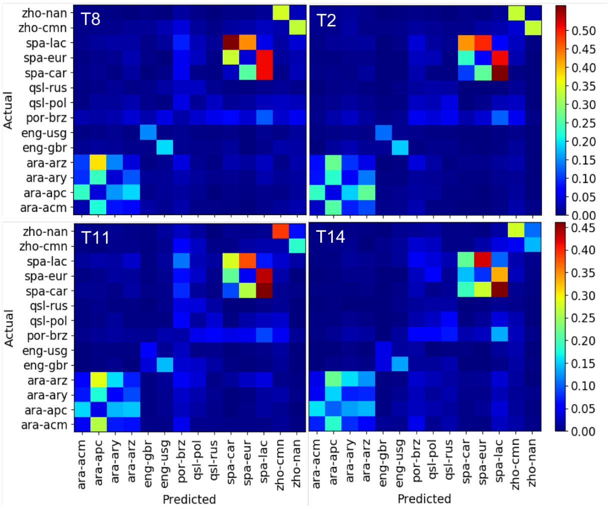

Figure 8: *Performance confusion matrices for the leading systems in LRE17.*

Figure 8 shows language recognition performance in form of confusion matrices for the 4 leading systems. Here, per language false-reject and false-accept rates are shown on diagonal and off diagonal of the matrices, respectively. The error rates are obtained for a target language prior probability of 0.5. Several important observations can be made from this figure. First, except for the Slavic cluster that has the lowest error rates, the language clusters are visibly highlighted. The languages within the Iberian cluster seem to be most confusable, followed by the Chinese and Arabic languages. Second, for `por-brz` majority of rows and columns are highlighted indicating that the systems tend to output a high likelihood score for Brazilian Portuguese language model irrespective of the actual language in test segments and vice versa. A similar behavior is also observed for `qsl-pol` language. Finally, `zho-nan` column is highlighted for the top performing team, but not for others.

Figure 9 shows a cross-year (LRE15 versus LRE17) performance comparison of the top-4 systems in terms of LRE15 performance metric, i.e., the average actual cost computed for the individual language clusters. We only included the MLS14 test segments for this analysis, because there are no AfV segments in LRE15. We note here that the language clusters differed across the two evaluations, and LRE15 included more languages (i.e., 20 in total). General performance improvements are observed for LRE17 systems (in particular for English and Chinese clusters), however there is a notable performance degradation for the Iberian cluster.

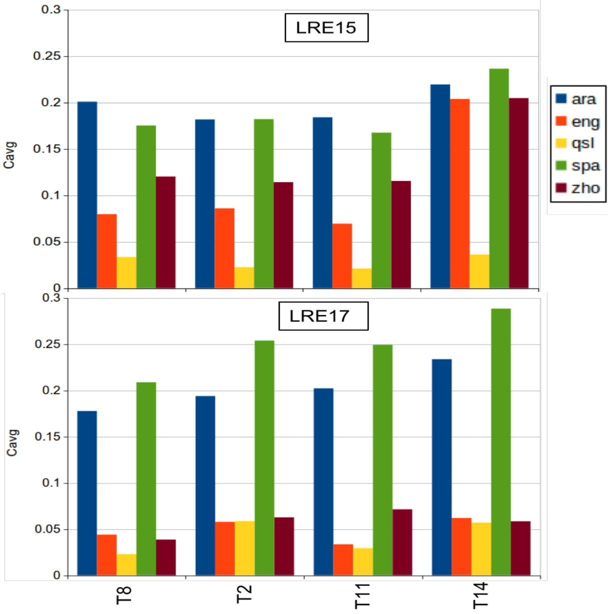

Figure 9: *Cross-year (LRE15 vs LRE17) performance comparison of the leading systems.*

5. Conclusions

We presented an analysis of system performance for LRE17 submissions, broken down by multiple factors such as data source and gender, as well as a cross-year (i.e., LRE15 versus LRE17) performance comparison of the top performing systems. It was observed that the difference in performance for MLS14 vs VAST or for male vs female was highly dependent on the underlying target language of the segments. We also compared performance of "single" best systems against the primary submissions in LRE17. While we saw notable benefit from fusion, strong "single" best systems proved to be as good.

6. 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.

7. References

[1] NIST, “NIST Language Recognition Evaluation,” https://www.nist.gov/itl/iad/mig/language-recognition, [Online; accessed 01-March-2018].

[2] 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 01-March-2018].

[3] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, “The 2016 NIST speaker recognition evaluation,” in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 1353–1357.

[4] NIST, “2017 Language Recognition Evaluation,” https://www.nist.gov/itl/iad/mig/nist-2017-language-recognition-evaluation, 2017, [Online; accessed 01-March-2018].

[5] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, “The 2017 NIST language recognition evaluation,” in *Proc. Odyssey*, Les Sables dÓlonne, France, June 2018.

[6] SIL International, “Documentation for ISO 639 identifier,” http://www-01.sil.org/iso639-3/, 2017, [Online; accessed 01-March-2018].

[7] 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 01-March-2018].

[8] 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 01-March-2018].

[9] 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 01-March-2018].

[10] 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 01-March-2018].

[11] J. Godfrey and E. Holliman, “Switchboard-1 Release 2,” https://catalog.ldc.upenn.edu/LDC97S62, 1993, [Online; accessed 01-March-2018].

[12] D. Graff, A. Canavan, and G. Zipperlen, “Switchboard-2 Phase I,” https://catalog.ldc.upenn.edu/LDC98S75, 1998, [Online; accessed 01-March-2018].

[13] D. Graff, K. Walker, and A. Canavan, “Switchboard-2 Phase II,” https://catalog.ldc.upenn.edu/LDC99S79, 1999, [Online; accessed 01-March-2018].

[14] D. Graff, D. Miller, and K. Walker, “Switchboard-2 Phase III,” https://catalog.ldc.upenn.edu/LDC2002S06, 2002, [Online; accessed 01-March-2018].

[15] D. Graff, K. Walker, and D. Miller, “Switchboard Cellular Part 1 Audio,” https://catalog.ldc.upenn.edu/LDC2001S13, 2001, [Online; accessed 01-March-2018].

[16] D. Graff, K. Walker, and D. Miller, “Switchboard Cellular Part 2 Audio,” https://catalog.ldc.upenn.edu/LDC2004S07, 2004, [Online; accessed 01-March-2018].

[17] M. P. Harper, “Data resources to support the Babel program,” https://goo.gl/9aq958, [Online; accessed 01-March-2018].

[18] 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.

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