Summary of the 2015 NIST Language Recognition i-Vector Machine Learning Challenge

Audrey Tong[1], *Craig Greenberg*[1], *Alvin Martin*[1], *Désiré Bansé*[1*], *John Howard*[1*], *Hui Zhao*[1*], *George Doddington*[1], *Daniel Garcia-Romero*[2], *Alan McCree*[2], *Douglas Reynolds*[3], *Elliot Singer*[3], *Jaime Hernández-Cordero*[4], *Lisa Mason*[4]

[1]National Institute of Standards and Technology, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, USA
[3]MIT Lincoln Laboratory, USA
[4]U.S. Department of Defense, USA
[*]Contractor

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

Abstract

In 2015 NIST coordinated the first language recognition evaluation (LRE) that used i-vectors as input, with the goals of attracting researchers outside of the speech processing community to tackle the language recognition problem, exploring new ideas in machine learning for use in language recognition, and improving recognition accuracy. The Language Recognition i-Vector Machine Learning Challenge, taking place over a period of four months, was well-received with 56 participants from 44 unique sites and over 3700 submissions, surpassing the participation levels of all previous traditional track LREs. The results of 46 of the 56 participants were better than the provided baseline system, with the best system achieving approximately 55% relative improvement over the baseline.

1. Introduction

The technology that automatically determines the language spoken in a given speech segment is often referred to as language recognition. NIST has coordinated evaluations to support research and advance state-of-the-art language recognition technology since 1996 [1]. In recent years one key advance in the field has been the use of i-vectors [2] to represent speech, as utilized in many systems with the most competitive performance.

In 2015 NIST coordinated its first language recognition evaluation using i-vectors as input. Modeled after the successful 2013-2014 Speaker Recognition i-Vector Machine Learning Challenges, the Language Recognition i-Vector Machine Learning Challenge had the goals of attracting researchers outside of the speech processing community to tackle the language recognition problem, exploring new ideas in machine learning for use in language recognition, and improving the performance of the technology, in particular for narrowband speech [3].

The Language Recognition Challenge defined the task as language identification: given a test segment and a set of possible target languages, determine which, if any, of the given target languages was spoken in the segment. In the evaluation plan for the language recognition challenge [4], the task was posed as open set, where the actual language spoken in the test segment could be a language not in the listed set of target languages, and introduced some degree of language confusability. The Challenge took place over a period of four months from May 1 to September 1, 2015.

We describe the data used in the Challenge in section 2 and the performance measurement in section 3. We summarize the participation and briefly mention the approach used by the system with the best performance observed in section 4. We discuss differences with the traditional track LREs in section 5. In section 6, we report the results and some observations and conclude in section 7.

2. Data

The data used in the Language Recognition i-Vector Challenge (i-vector LRE) were speech data recordings collected over telephone channels (telephone conversations and narrowband broadcasts) and came from previous NIST Language Recognition Evaluations spanning from 1996 to 2011 and from selected sources developed for other programs (e.g., the IARPA Babel Program [5]). The Challenge introduced some degree of confusability between the in-set and out-of-set languages and among the in-set languages themselves. This was accomplished using a data driven approach described below.

To reduce possible source to language effect, only languages that were represented in several sources were selected. This yielded a total of 65 viable languages from which to choose. This data pool was partitioned into three datasets: *training*, *development*, and *test*. To create the partitions, for each of the 65 languages, 100 random recordings/files were selected for the test set, then another 100 files (or the remainder if less than 100) for the development set, and finally another 300 files (or remainder if less than 300) for the training set after the requirements for the test and development sets were met. This algorithm produced 65 languages for test and development and 59 languages for training. The six languages included in test/development but not in training were designated as out-of-set. Furthermore, a confusion matrix was computed for the 59 languages. The pairs of languages were ordered from most to least confusable pairs. Every other pair was selected, with one language in the pair going into the out-of-set group unless that

language was already in the group. This yielded an additional nine languages for the out-of-set for a total of 15, reducing the training set to 50 languages designated as target languages.

Language labels were given with the training set for the 50 target languages. However, no language labels were given for the development set. The development set intended to provide an additional resource for systems to learn out-of-set languages through unsupervised clustering. The test set was further divided into two subsets called *progress* and *evaluation* with a 30%/70% sub-division. The subdivision was a random selection stratified by language. The purpose of the subdivision in the test set was to provide score feedback on a portion of the test set (progress) and keep the rest (evaluation) blind to assess whether systems overfitted to the progress set. Table 1 summarizes the data statistics across the three datasets.

Table 1: *The size of the datasets used in the Language Recognition i-Vector Challenge. Some languages in the Development set did not have 100 files due to data availability.*

Set	Num. of Languages	Num. of Segments per Language	Total Segments
Train	50	300	1500
Dev.	65 • 50 target • 15 out-of-set	$\approx$ 100	6431
Test	65 • 50 target • 15 out-of-set	100	6500 • 5000 target – 1500 progress – 3500 evaluation • 1500 out-of-set – 450 progress – 1050 evaluation

A single segment was extracted from each selected recording/file. The speech duration of the segments followed a log-normal distribution with a mean of 35 seconds. These segments were then converted to i-vectors. The i-vectors, along with the segment speech durations, were given to the participants for processing unlike the traditional track LRE, which supplies the actual audio. The i-vectors were 400 dimensional vectors produced by a system developed by the Johns Hopkins University Human Language Technology Center of Excellence and MIT Lincoln Laboratory. More information about the i-vector system can be found in [2]. See Figure 1 for the distributions of the segment durations for the three sets. The training and test sets contained relatively more shorter duration segments than the development set.

While it is beyond the scope of this paper to describe all of the sources used in fine detail, it is worth mentioning some distinguishing characteristics among the sources. As described in [6], the LDC CallFriend corpus contained 5-30 minute telephone conversations where the speakers, conversing in their native language, knew each other and lived in the same country, typically the U.S., Canada, Puerto Rico, or Dominican Republic, while the LDC CallHome corpus contained up to 30 minutes telephone conversations. The Fisher corpus from LDC contained 5-minute telephone conversations between two strangers on assigned topics, resulting in a more diverse speaker population as described in [7]. The LDC Mixer corpus was similar to the Fisher corpus except that the speakers were bilingual, conducting some conversations in their native language and some in their second language; additionally, the conversations were recorded over multiple microphones as described in [8]. The OHSU corpus[1], collected by the Oregon Health & Science University, contained telephone conversations of people recruited from various locations in the U.S. to make calls to people they knew, half of whom resided in the U.S. while the other half resided in other countries. The Babel corpus, collected by Appen, contained telephone conversations recorded in-country for the languages of interest using a variety of handsets (e.g., cell phone, table microphone) and recording conditions (e.g., on the street, in a busy café) as described in [9]. LRE07 contained unused data from CallFriend, Mixer, and OHSU [1]. The NBBN corpus contained narrowband broadcast speech from MIT Lincoln Laboratory. LRE09 contained a mix of telephone conversations from previously mentioned corpora and narrowband broadcasts from Voice of America in multiple languages. Finally, the LDC LRE11 corpus contained both telephone conversations and narrowband broadcasts using similar protocols to those used in CallFriend and VOA as stated in [10].

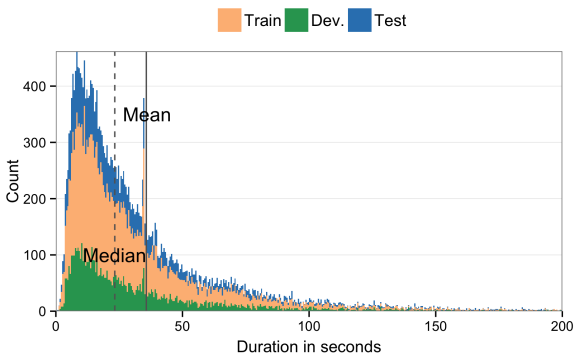

Figure 1: *Distributions of the segment speech durations for the training, development, and test datasets. The graph displays only to 200 seconds but segment durations extend beyond 800 seconds. Solid line indicates the overall mean duration and dashed line indicate the overall median duration.*

As stated above, because the data came from a variety of sources collected by different organizations and under slightly different conditions, only languages that were present in several sources were selected to minimize the source to language effect. We note the distributions of sources across languages are similar across the three datasets as shown in Figure 2.

3. Performance Measurement

The metric used to compute the system performance and ranking is the cost function defined as a weighted sum of the error

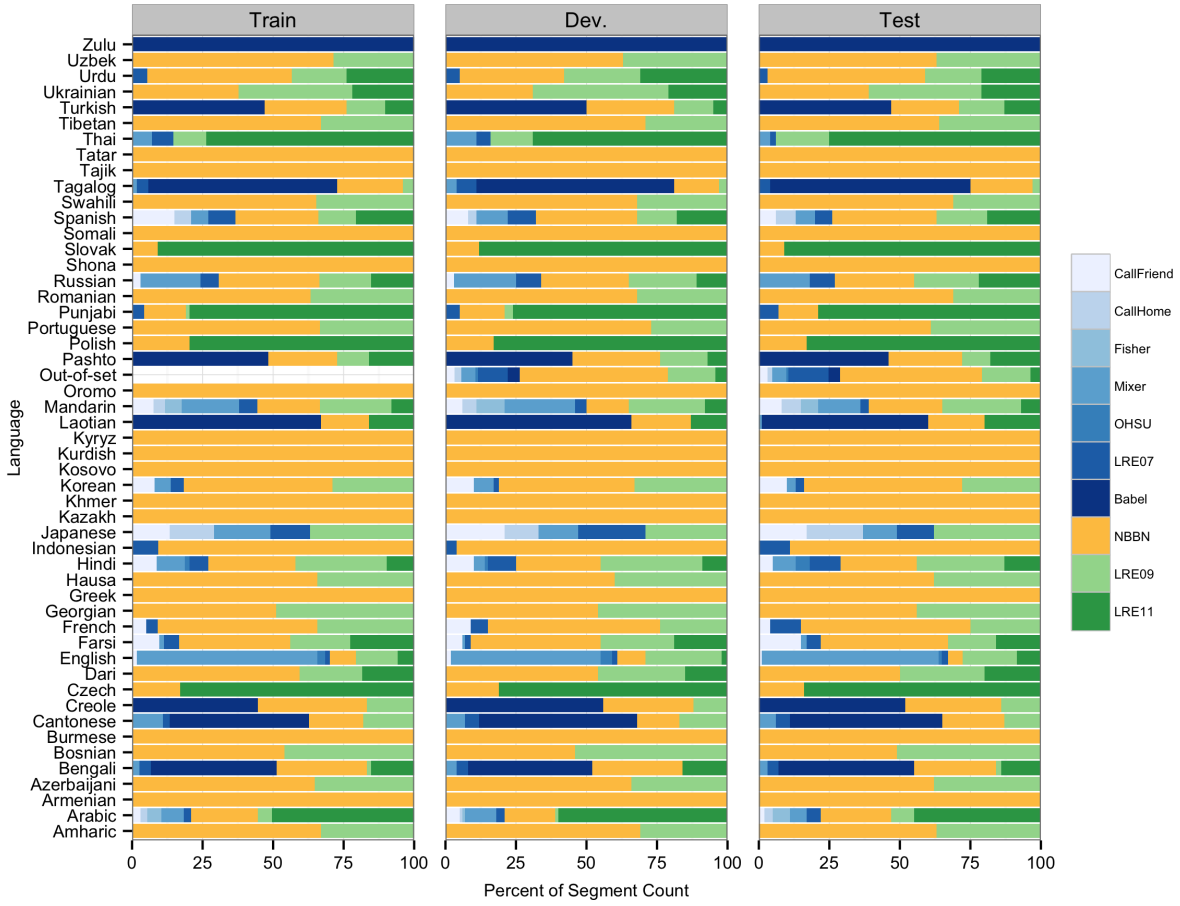

Figure 2: *The distribution of the segments across target and out-of-set languages and corpora for the three datasets (training, development, and test). Shades of blue, yellow, and green indicate the data is telephone conversations, narrowband broadcasts, and a mix of both, respectively. Note there is no data for the out-of-set language in the training set.*

rate for each target language and the error rate for segments not in the target language set, which is:

$$Cost = \frac{(1 - P_{oos})}{n} \Sigma_k^n P_{error}(k) + P_{oos} P_{error}(oos) \quad (1)$$

where $P_{error} = \frac{\#errors}{\#trials}$ for target language class k and for out-of-set class oos, $n = 50$, and $P_{oos} = 0.23$.

During the Challenge period, the scores for the progress set were computed as feedback to the participants. The best score achieved on the progress set determined the final submission for each participant. The overall ranking was determined by the scores on the evaluation subset of the final submissions.

4. Participation

We saw increased participation in this i-vector Challenge as compared to traditional LREs (see Figure 3) in terms of the number of submissions and participating sites. A total of 136 participants from 31 countries registered, of which 78 downloaded the evaluation data and 56 submitted a total of 3773 valid submissions. The 56 who participated came from 44 unique sites/organizations, suggesting some sites had multiple registrations. The Challenge made no distinction between participants and sites and counted each registration as a single identity allowing members of the same site to compare their methods against each other as well as against members of other sites. For clarity, this paper refers a participant as a single registration regardless of his/her site affiliation.

The Challenge took place over a period of four months in 2015 from May 1 to September 1. While the Challenge has ended, submissions can still be made to the system for scoring for the foreseeable future, but participant rankings will not be updated. As of January 19, 2016, 4021 submissions have been made to the scoring server.

The system with the best performance observed on the evaluation subset employed a fusion approach that combined multiple sub-systems based on commonly used techniques in language recognition including linear discriminant analysis, support vector machines, multi-layer perceptrons, deep neural networks, and multi-class logistic regression. However, this system introduced two new methods: (1) one that enhanced out-of-set detection by exploiting the out-of-set distribution from the given unlabeled development set and (2) one that transformed the given training data to the score space with respect to all other data points in the training set. More information about this system can be found in [11].

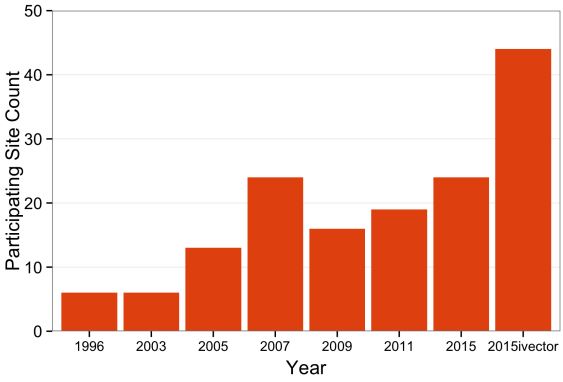

Figure 3: *Number of participating sites from all NIST-run language recognition evaluations.*

5. Contrast with Traditional NIST LREs

The i-vector LRE was quite different from the traditional audio-based LREs that NIST runs, including the most recent traditional LRE in 2015 (LRE15). The major defining difference was the input to the system, using i-vector representations instead of actual speech segments. The key idea was that by eliminating the front-end speech processing the complexity was reduced, facilitating people from outside the speech processing community to tackle the language recognition problem. Additionally, by having a fixed front-end, one can better understand the algorithmic differences among the different back-ends.

Another difference was in the task definition. The task for the i-vector LRE was language identification instead of language detection. The language identification task was an N-class problem where the system was given the test segment and a list of possible languages and was asked to determine which language in the list was spoken in the segment, whereas the language detection was a two-class problem where the system was given the test segment and a target language and was asked to determine if the target language was spoken in the test segment. The metric used to compute the system performance in the i-vector LRE was also different using a cost based on error rates while the traditional LREs used a cost based on miss and false alarm rates.

The number of target languages was also different, with the i-vector LRE having 50 languages, while the traditional LREs included 10 to 25 languages. Unlike the prior LREs, the i-vector LRE had labeled out-of-set languages as well as unlabeled development sets. For the i-vector LRE, only one segment was taken from each original recording instead of multiple segments, and the segment durations followed a log normal distribution instead of a uniform distribution.

The duration of the evaluation was dissimilar. The i-vector LRE took place over four months and allowed multiple submissions with feedback score provided for a small portion of the test set (e.g., the progress subset) after each submission, enabling participants to make incremental improvement over the duration of the Challenge, while the rest of the test set (e.g., the evaluation subset) remained hidden to gauge the true performance on completely unseen data. The traditional evaluation usually took place over a two-week period with no score feedback for any portion of the test set. Another difference was in the way the i-vector LRE was conducted in that participants conducted all evaluation related activities via a web-based platform. The level of participation in the i-vector LRE was much higher, surpassing all the prior LREs as noted in section 4.

6. Results

In this section we report the results obtained from the Challenge and also note some observations. Figure 4 shows the overall best results (lowest Cost) achieved by each participant on the evaluation subset on the best submission as determined by the lowest score received on the progress subset. As comparison points, the results of a baseline, oracle, and do-nothing system are also shown. The simple *baseline* system modeled each target language as the average of the training i-vectors and used a similarity function (cosine distance) to predict the language of the test i-vectors. The *oracle* system was the same as the baseline system except that the oracle system had access to the language label of the development set. The *do-nothing* system did nothing except using the knowledge given in the evaluation plan the prior probability given for out-of-set P_{oos} and classifying all segments as out-of-set. Forty-six of the 56 participants received better scores than the baseline system, with the top system achieving 54.56% relative improvement over the baseline score of 0.39 on the evaluation subset. While the baseline system was not a state-of-the-art system, it represented a reasonable point to compare the submissions.

The oracle system had knowledge about the out-of-set i-vectors, and so was used to represent the value of the labels. Six systems were able to beat the oracle system, with the top system achieving a 10.83% relative improvement over the oracle score of 0.19 on the evaluation subset. Note also that all systems except one had better scores than the do-nothing system suggesting that most of the participant systems were non-trivial.

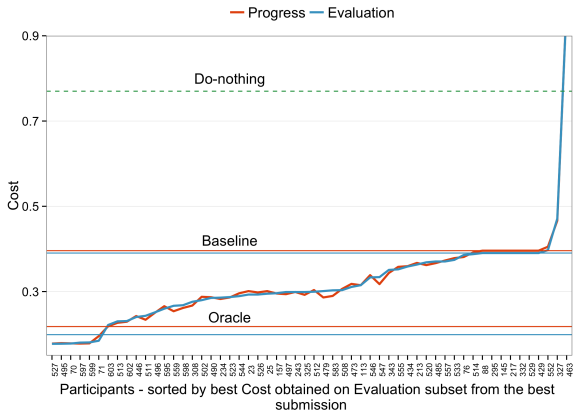

Figure 4: *Lowest Cost achieved by each participant on the Evaluation subset of the best submission. The best submission was determined by the lowest Cost achieved on the Progress subset. The y-axis displays only to 0.5, but the cost for the last point extends to almost 1.0.*

Figure 5 shows the number of submissions received from each participant, using the same ordering as that of Figure 4. Figure 6 shows the best results each day on the progress subset for the duration of the Challenge. Improvement was seen throughout the Challenge. Figure 7 shows the error rate for

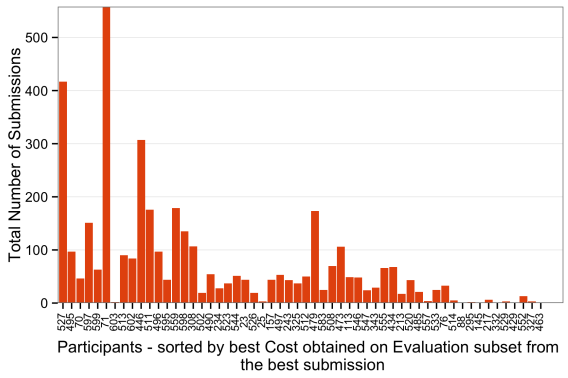

Figure 5: *Total number of submissions received from each participant, ordered by best result achieved on the Evaluation subset of the best submission. The best submission was determined by the lowest Cost achieved on the Progress subset.*

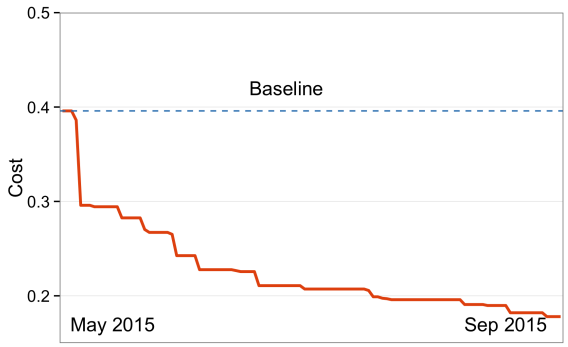

Figure 6: *Daily lowest Cost obtained on the Progress subset during the official period of the Challenge.*

each target language including the out-of-set category averaged across the best systems from the top five participants. The results suggest some languages are harder to recognize than others. However, these results did not factor in the channel effect as each language had very different data composition in terms of channel variations as shown in Figure 2. Figure 8 shows the error rate for each duration bin, also averaged across the best systems from the top five participants. We note, not surprisingly, higher error rates on shorter segments.

7. Conclusions

We presented a summary of the first NIST Language Recognition i-Vector Machine Learning Challenge. The objective of the Challenge was to advance the state-of-the-art in language recognition technology by providing a new paradigm designed to attract researchers outside of the speech community to solve the language recognition problem, to investigate new methods in machine learning for use in language recognition, and to improve the performance of the technology. The Challenge was well-received with more participating sites than any other NIST-run LREs. Some of the participating sites in the Challenge had never participated in previous NIST-run LREs suggesting the

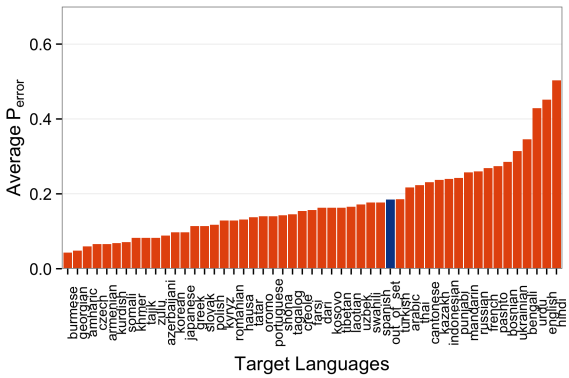

Figure 7: *The error rate of each target language and out-of-set category averaged across the top five systems.*

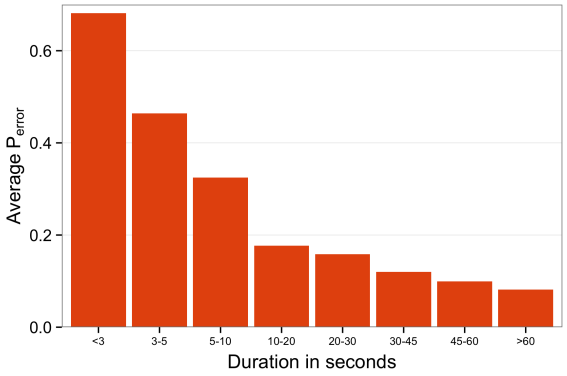

Figure 8: *The error rate for each duration bin averaged across the top five systems.*

paradigm attracted a wider group of researchers. Even after the Challenge ended in September 2015, submissions are still being received to date. Unfortunately, not all participants submitted system descriptions so it was not clear whether those participants investigated new methods. However, we did learn that the system with the best performance observed on the evaluation subset developed a novel technique to improve out-of-set detection and introduce a new kernel mapping method for use with the support vector machine. We compared the systems in the Challenge against a simple baseline as well as an oracle system that had knowledge of the out-of-set language labels. Six systems achieved better scores than the oracle scores suggesting approaching state-of-the-art.

The current plan is to keep the web-based scoring platform up for the foreseeable future as a system development tool for researchers who work on this problem while plans for the next challenge are being considered. Researchers can go to http://ivectorchallenge.nist.gov to download i-vector data for training and development as well as benchmark their systems on the test data.

8. References

[1] Alvin F. Martin, Craig S. Greenberg, John M. Howard, George R. Doddington, and John J. Godfrey, "NIST

language recognition evaluation past and future," in *Odyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland, June 2014, pp. 145–151.

[2] Najim Dehak, Pedro A. Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak, "Language recognition via ivectors and dimensionality reduction," in *Interspeech*, Florence, Italy, August 2011, pp. 857–860.

[3] "i-vector machine learning challenge," Online: http://www.nist.gov/itl/iad/mig/ivec.cfm, Last access: March 30, 2016.

[4] "The 2015 language recognition i-vector machine learning challenge," Online: http://www.nist.gov/itl/iad/mig/upload/lre_ivectorchallenge_rel_v2.pdf, Last access: March 30, 2016.

[5] "Babel," Online: http://www.iarpa.gov/index.php/research-programs/babel, Last access: March 30, 2016.

[6] Mark Liberman and Christopher Cieri, "The creation, distribution and use of linguistic data," in *LREC*, Grenada, Spain, May 1998.

[7] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *LREC*, Lisbon, Portugal, May 2004, pp. 69–71.

[8] Christopher Cieri, Joseph P. Campbell, Hirotaka Nakasone, David Miller, and Kevin Walker, "The mixer corpus of multilingual, multichannel speaker recognition data," in *LREC*, Lisbon, Portugal, May 2004, pp. 627–630.

[9] Mary P. Harper, "Data resources to support the babel program intelligence advanced research projects activity (iarpa)," Online: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/harper.pdf, Last access: March 30, 2016.

[10] Stephanie Strassel, Kevin Walker, Karen Jones, Dave Graff, and Christopher Cieri, "New resources for recognition of confusable linguistic varieties: The lre11 corpus," in *Odyssey: The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 202–208.

[11] Hanwu Sun, Trung Hieu Nguyen, Guangsen Wang, Kong Aik Lee, Bin Ma, and Haizhou Li, "I2r submission to the 2015 nist language recognition i-vector challenge," February 2015.