INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The MIT-LL, JHU and LRDE NIST 2016 Speaker Recognition Evaluation System

Pedro A. Torres-Carrasquillo[1], *Fred Richardson*[1], *Shahan Nercessian*[1], *Douglas Sturim*[1], *William Campbell*[1], *Youngjune Gwon*[1], *Swaroop Vattam*[1], *Najim Dehak*[2], *Harish Mallidi*[2], *Phani Sankar Nidadavolu*[2], *Ruizhi Li*[2], *Reda Dehak*[3]

[1]MIT Lincoln Laboratory, Lexington, MA, USA
[2]Johns Hopkins University, Baltimore, MD, USA
[3]LRDE-EPITA, Paris,France

{ptorres,frichard,Shahan.Nercessian,sturim,wcampbell,gyj,Swaroop.Vattam,}@ll.mit.edu,
{ndehak3,rli33}@jhu.edu,{mallidi.harish,phanisankar.nsv}@gmail.com,
reda.dehak@lrde.epita.fr

Abstract

In this paper, the NIST 2016 SRE system that resulted from the collaboration between MIT Lincoln Laboratory and the team at Johns Hopkins University is presented. The submissions for the 2016 evaluation consisted of three fixed condition submissions and a single system open condition submission. The primary submission on the fixed (and core) condition resulted in an actual DCF of .618. Details of the submissions are discussed along with some discussion and observations of the 2016 evaluation campaign.

Index Terms: speaker recognition, speaker evaluation.

1. Introduction

Speaker recognition, in loose terms, is the process of associating a speech utterance whose speaker's identity is unknown with another utterance whose speaker's identity is known. Over the last few decades the National Institute of Standards and Technology (NIST), located in the United States, has conducted evaluations of speaker recognition technology to assess the performance of speaker recognition systems developed by interested parties across the world. NIST conducted its most recent speaker recognition evaluation (SRE) [1] during 2016 and this evaluation represents a shift in paradigm over previous evaluations.

The 2016 SRE included a number of new challenges for system developers that had not been presented in previous evaluations. The highlights of these new challenges included:

- Focus on data collected outside of the North American telephone network
- No English data usage (evaluation data languages used were Cantonese and Tagalog)
- Small labeled in-domain development set
- Availability of unlabeled pool of in-domain data for system development
- Duration limited to 10-60 seconds for the evaluation samples with enrollment limited to 60 seconds
- Duration limited to 10-60 seconds for the evaluation samples with enrollment limited to 60 seconds
- Fixed training condition as core evaluation condition

In this paper the submission that resulted from the collaboration between MIT Lincoln Laboratory and Johns Hopkins University to the 2016 SRE is described. The submission included 3 submissions for the fixed training condition, consisting of a fusion of multiple systems, and a single system submission to the open training condition.

2. Data Description

As mentioned in the previous section the 2016 NIST SRE featured a new core (required) condition that featured a "fixed" or closed training condition. The closed condition requirement constrained all system training to data made available by the Linguistic Data Consortium [2] and more specifically only included data available to participants from previous SRE campaigns. The motivation for this condition is likely very clear and implies that NIST and the community can focus on evaluating and understanding the algorithms used by the system developers instead of convolving the interaction between the algorithms and the data engineering aspects of system development. Although the MIT-LL/JHU/LRDE team entered a submission to the open condition, during the development stage all decisions were made focusing on the fixed training condition.

The team made the decision to partition the data in a way that closely resembles the expected conditions of the evaluation data and in particular in a way that results in a similar mismatch to that likely to arise at evaluation time. The decision was made to construct two sets, one for English only and a second one for non-English languages only. The interaction between these two sets was expected to address the language mismatch between the development data available which was mainly English data and the data in the evaluation which was non-English (Cantonese and Tagalog).

In more detail, the data partitions were constructed as follows. The English partition included data from Switchboard (SWB) 1 and 2 [3] plus data from previous NIST SREs including 2004-2010. The non-English partition included utterances from SREs 2004-2008. For each the English and the non-English sets, speakers with at least ten sessions and 120 second minimum duration were selected. From this pool and for each partition, 3

This work was sponsored by the Department of Defense under Air Force contract F19628-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-537

sessions were used for enrollment and the remaining sessions were used for the testing phase of system development. The testing cuts were processed to mimic the expected uniform duration distribution for the evaluation set in the 10-60 second range. Additionally, the labeled in-domain development set (SRE 2016 dev) distributed by NIST was left intact and used for additional diagnostics and as part of the final submission decision. The SRE 2016 dev set consisted of 20 speakers (10 for Mandarin and 10 for Cebuano), evenly distributed in terms of gender. Table 1 shows the number of speakers used for the English and Non-English partitions with its gender split.

Table 1. *Speaker and language distribution for the available data used for the development partitions*

Data Partition	Languages	Speakers Available
English	English	254 (109 m/ 146 f)
Non-English	Mandarin	16 (11 m / 5 f)
	Russian	14 (5 m / 9 f)
	Thai	18 (4 m / 14 f)
	Yue	21 (5 m / 16 f)
	Arabic	19 (10 m / 9 f)

3. System Setup

The Lincoln-JHU-LRDE team considered various candidate systems. Most of the systems submitted were based on i-vector approaches [4]. In this section the general framework and decisions for the system components training is described.

3.1. Common i-vector component training

The general i-vector framework requires training a number of common components. These components include the alignment component (be it a UBM [5] or a DNN-based alignment [6]), the T-matrix, the mean centralization and whitener stage, and the PLDA scoring stage.

For the systems, the data partition was used for training as follows. The UBM and T-matrix stages used full duration Switchboard cuts. The whitener stage used in-domain data based on the task under consideration. For example, during the development stage the Non-English data set was used for the Non-English task while the unlabeled SRE 2016 Dev set was used for the SRE 2016 task. In the case of the JHU/LRDE systems, the system components used the Switchboard dataset, and all the hyper-parameter partitions of the English and non-English sets described above were used to train the UBM and T-matrix with the same sets as the Lincoln systems used for whitening [7] and PLDA training.

3.2. Voice Activity Detection

Although the feature processing for the Lincoln and JHU/LRDE systems is not homogenous either across or within sites, the VAD used was uniform within two varieties. The VAD alternatives included a GMM based approach trained exclusively on a small pool of SWB utterances. The GMM used 128 mixtures and the output further processed by an energy based detector. The threshold of the energy detector was increased to reduce the amount of speech left for processing on the SRE 2016 development set. Some of the JHU/LRDE systems in the contrastive submissions also considered VAD using Kaldi [8].

3.3. System Submissions

The team submission included three systems to the fixed training condition featuring a primary submission and two contrastive submissions as described below.

3.3.1. Primary Submission

The primary submission for the MIT-LL/JHU/LRDE team consisted of a fusion of four systems. The four systems submitted included are described below with the fusion strategy, score normalization and adaptation technique used included after the core system descriptions.

3.3.1.1 JHU/LRDE MFCC and pitch i-vector system

The JHU/LRDE MFCC i-vector system used a concatenation of two feature sets. The first stream consists of a 60-dimensional feature vector that included 20 static coefficients along with first and second derivatives. Cepstral mean subtraction on a 300ms sliding window was used. The second "stream" is based on pitch features. These features were extracted using the Kaldi pitch extractor [9] produces a two dimensional output: a normalized cross correlation function (NCCF) and the pitch (in Hz). NCCF values range between -1 and +1. NCCF is higher for voiced frames. These two features are further processed to obtain three-dimensional features (pov-feature, pitch-feature and delta-pitch-feature). Pov-feature is a warped version of the NCCF, the pitch-feature is a log-pitch with POV-weighted mean subtraction over 1.5s window, and the delta-pitch-feature is a delta feature computed on raw log pitch. The augmented feature vector is then processed by a conventional i-vector system parameterized by a 2048-mixture UBM using a full covariance and a 600-dimensional i-vector.

3.3.1.2 MIT-LL MFCC i-vector system

The MIT-LL i-vector system uses a 40-dimensional feature vector including 20 MFCC static coefficients its derivatives. The obtained feature vector is then processed through the i-vector system featuring a 2048-mixture UBM with diagonal covariance and 600-dimensional i-vector.

3.3.1.3 MIT-LL tandem SDC BNF i-vector system

The MIT-LL tandem shifted delta cepstral (SDC) bottleneck features (BNF) system is an i-vector system combining features from an SDC feature stream and a BNF stream [10]. The SDC feature set [11] uses the conventional 7-1-3-7 parameterization along with the 7 static coefficients. The SDC stream is then concatenated with an additional set of 80 features derived from a DNN via Bottleneck features. The DNN used was trained using SWB-1 English. For this system all hyperparameters were trained as outlined above with a minor modification for the PLDA stage where only speakers with a minimum of 4 cuts were used.

3.3.1.4 MIT-LL Denoising stats i-vector system

The MIT-LL denoising stats i-vector system used impulse responses estimated from Mixer 2 telephone / microphone sessions and applied to the SRE closed set data (specifically the Mixer 1&2 studio LDC sessions). The input to the DNN is a 21 frame window using a set of stacked 40-dimensional MFCC (as in the Section 3.3.1.2) and the target was the 40 dimensional feature vector at the center of the input window extracted from the clean data. The general setup for this system is very close to that described in [12].

A 7x1024 layer senone classifying DNN was trained on 300 hours of SWB with ~8K senone targets. The first four layers were the same as the first four layers of the denoising DNN described above and only the last three layers were trained to minimize cross entropy of the senone posterior output prediction. The configuration is otherwise similar to that described in [10].

The set of 8K senones were clustered by modeling each senone us an i-vector, projecting down using an LDA transform and clustering via K-means to obtain 2048 senone clusters. Silence was not clustered but the silence labels were preserved and assigned unique cluster labels. A senone cluster classifying DNN was then trained using the same features and architecture described in the previous paragraph along with the new frame-level senone cluster labels.

The final denoising stats system used the senone cluster classifying DNN posteriors together with the 40 baseline MFCC features to create super vectors. The denoising stats system super vectors were used to estimate a T-matrix and then to extract the final i-vectors.

3.3.1.5 Domain Adaptation

Each of the systems included used a system adaptation scheme that included two components. First, each of the systems used the available NIST SRE 2016 in-domain data (both labeled and unlabeled) for whitening and mean centralization. Second, every system used a multi-stage PLDA adaptation technique. Each of the PLDA matrices was initially trained on the development English partition and MAP-adapted to the non-English task using a weight adaptation of $\alpha = 0.5$ [13]. The non-English matrices were then adapted to the in-domain data set using a more conservative weight of $\alpha = 0.2$. The adaptation values were obtained by sweeping across the range of 0-1 but also considering the amount of data available for each task. Additionally, the NIST SRE 2016 development set was used using both the labeled and the unlabeled partitions. The labeled partition was used as provided by NIST. The unlabeled partition was clustered and combined with the labeled partition. The clusters for the unlabeled data set were obtained by applying agglomerative hierarchical clustering to the whitened i-vectors. Outlier rejection was applied by filtering out clusters with fewer than 3 speakers and by computing z-scores and thresholding those scores. From the clustering process 80 clusters were inferred (70 cluster for the major languages and 10 clusters for the minor languages).

3.3.1.6 Score Normalization

Score normalization of the obtained scores was conducted using an adaptive z-norm approach [14]. The adaptive z-norm approach used the top 200 scoring utterances from the NIST SRE 2016 data set (both labeled and unlabeled) against each of the available models.

3.3.1.7 Calibration and System Combination

Calibration of the system scores was performed by a using a logistic regression approach and cross validation. For each system, the regression was trained on non-English partitions and the 2016 dev set and applied to the evaluation data set. The final composition of the primary submission was based on the performance of the systems on two tasks: the non-English development task and the SRE 2016 DEV task as defined by NIST.

3.3.2. Secondary Submissions

The team also submitted to secondary or contrastive submissions. Although these are not described in detail in this paper these submission included two additional systems that were small variations of the systems described above plus a new system based on sparse coding adapted from [15].

3.3.3. Open Submission

The team decided to submit a single system for the open training condition consisting of a multilingual bottleneck features system that feature training data from the Babel language set.

3.3.3.1 Multilingual bottleneck (MLB) features using multistream processing i-vector system

The MLB feature i-vector system uses a 47-dimensional MFCC vector and a 3-dimensional set of pitch features. These features are grouped into 6 streams. These streams are made up of 5 sub-band streams from the MFCC features and 1 pitch stream. The 6 streams are provided as input to a Stream Dropout neural network. The Stream Dropout network consisted of 2 stages. The first stage included 6 networks, one for each stream, to extract stream specific bottleneck features. Each of the first stage networks consists of two 1500 rectified linear units (RELU) followed by a 40 dimensional linear layer. The output bottleneck features in each stream are then concatenated and provided as input to a second stage, a fusion network. This second stage consisted of two 1500 RELU layers, followed by a 60 dimensional bottleneck layer, and two more 1500 RELU layers and a final block-softmax layer. The entire network was trained using a mini batch stochastic gradient descent approach, using data from 14 languages available from the IARPA Babel program. The final feature vector results in 60-dimensional set. The obtained features used as input to the i-vector system that featured a 2048 mixtures UBM with full covariance and i-vector dimension of 600.

4. Evaluation Results

This section presents the results obtained for all system submissions on the 2016 SRE and then discusses some initial analysis of the obtained results.

Table 2 presents the results obtained for each of the submissions of the MIT-LL/JHU/LRDE team.

From the results in Table 2 a few observations are in order. First, the calibration of the system is on target. This result is interesting as the expectation was that calibration, given the multiple unknowns on this evaluation, would likely be problematic. Also, it is shown that the performance of the primary submission was very close to the performance obtained on the contrastive submission although the second contrastive submission would have made a better choice. The last observation from this table is the performance observed on the open training condition submission. It is shown that the performance of the open submission is well below par when compared to the fixed condition submission. There are two issues that could account for this result. First, the fixed condition submissions are based on a fusion of multiple systems while the open submission was a single system and second, the lack of emphasis on the open condition during the development phase.

Table 2. *MIT-LL/JHU/LRDE evaluation systems performance in terms of act/min DCF*

System	SRE16 Eval
Primary	0.634 / 0.623
Secondary 1	0.644 / 0.636
Secondary 2	0.629 / 0.618
Open	0.729 / 0.693

4.1. Post-Evaluation Results

This section presents a number of results that the team obtained on the initial analysis conducted after the evaluation was completed. These results are the initial phase of what will likely be a long period of analysis to breakdown all the factors that affected the performance of systems in this evaluation.

4.1.1. Primary Submission Breakdown

Table 3 shows the performance of the primary submission as the number of systems is increased until all four systems have been fused. The performance observed is additive and not a single system performance, i.e., each row represents the performance of all the systems previously shown in that column. In general terms, two observations are in order. First, that fusion of systems does result in improved performance over a single system. However, it is also clear most of the performance gain obtained from fusing multiple systems is obtained after fusing the first two systems. Additionally, it is also shown that the three system fusion was slightly better than the performance obtained by the fusion of four systems.

Table 3. *Performance breakdown of MIT-LL/JHU/LRDE primary submission in terms of act/min DCF*

System	SRE16 Eval
JHU/LRDE MFCC and pitch	0.686 / 0.667
MIT-LL tandem SDC BNF	0.632 / 0.629
MIT-LL MFCC	0.627 / 0.620
MIT-LL Denoising stats	0.634 / 0.623

4.1.2. Effect of score normalization

As described earlier, the system submitted by the MIT-LL/JHU/LRDE team featured the use of an adaptive z-norm scheme for score normalization. Table 4 describes the effect of the z-normalization process as a function of the number of utterances used. The table shows that as long as z-norm was used the effect of the size of the pool of utterances used for scoring was of limited effect. However, additional insight is obtained when the effects of z-norm are observed in combination with the hyperparameter adaptation scheme. Table 5 shows the effects of combining the PLDA adaptation scheme with z-normalization.

Table 4. *Effect of number of utterances used for z-normalization in terms of act/min DCF*

# of utterances	SRE16 Eval
0 (No-norm)	0.738 / 0.728
50	0.671 / 0.669
100	0.661 / 0.658
200	0.670 / 0.651
400	0.698 / 0.650
800	0.725 / 0.651
1600	0.699 / 0.656
All	0.826 / 0.678

The results in Table 5 show that there were improvements obtained from both PLDA adaptation and adaptive z-normalization over the un-normalized alternative. It is also very interesting that most of the gains obtained were due to the adaptive z-norm and that PLDA normalization on its own was only helpful in terms of calibration. The combination of both PLDA and adaptive z-norm resulted in about a 10% gain compared to using adaptive z-norm alone.

Table 5. *Interaction between PLDA matrices adaptation and adaptive z-normalization in terms of act/min DCF*

PLDA Adaptation	Adaptive z-norm	SRE16 Eval (act/min DCF)
		6.73 / 0.959
✓		0.960 / 0.960
	✓	0.738 / 0.728
✓	✓	0.670 / 0.651

4.1.3. Open System Submission

As shown earlier in Section 4, the performance of the open training condition was below that of the fixed training condition. Table 6 shows an extension of the work originally completed for the open training submission. In the table results are shown for combining multiple systems instead of a single system submission. The performance of the system is very close to that obtained in the fixed training condition resulting in about a 5% improvement. One possible explanation for the small gains is that the extra data available for training is not representative of the evaluation domain data and therefore additional data has limited impact. It is also worth noting that for this fusion only the multilingual BNF system exploits the use of the additional data. At this point additional work is needed to further understand this result.

Table 6. *Effect of fusion of multiple systems*

System	SRE16 Eval (act/min DCF)
JHU/LRDE MFCC and pitch	0.686 / 0.667
MIT-LL tandem SDC BNF	0.632 / 0.629
JHU/LRDE multilingual BNF	0.609 / 0.606

5. Conclusions and Future Work

In this paper the MIT-LL/JHU/LRDE NIST SRE 2016 submission systems are described. The 2016 was extremely challenging as it represented a major paradigm shift within the context of recent NIST SREs. The submissions consisted of well-known i-vector systems and also included various schemes for using the available development data, featuring PLDA matrix adaptation and adaptive z-normalization. The adaptation scheme employed resulted in good performance and excellent calibration on the evaluation data.

The paradigm shift on this evaluation resulted in a large degradation in performance compared to the recent trends observed in these evaluations (close to an order of magnitude worse). There are a number of possible areas that require further analysis to determine the reasons for the performance degradation. These reasons include a number of mismatch conditions in the data including a new channel, new languages, collection platform and duration.

An additional area of future work includes understanding newer DNN techniques and why many of the more recent DNN approaches fail to provide good performance in this evaluation.

As a final word, the new evaluation paradigm has open up the door for the development of additional understanding in the speaker recognition community and it should reenergize the community over the next few years.

6. References

[1] 2016 Speaker Recognition Evaluation plan. https://www.nist.gov/file/325336

[2] https://www.ldc.upenn.edu/

[3] J. J. Godfrey, E. C. Holliman and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development", *Proceedings of the IEEE ICASSP*, San Francisco, CA, 1992, pp. 517-520

[4] N. Dehak et al., "Front-end factor analysis for speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] D. A. Reynolds, T. F., Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing Review Journal*, January 2000.

[6] Y. Lei, N. Scheffer, L. Ferrer and M. McLaren, "A Novel Scheme For Speaker Recognition Using A Phonetically-Aware", *Proceedings of the IEEE ICASSP*, Florence, Italy, 2014, pp. 1714–1718.

[7] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 249-252.

[8] D. Povey, et al. "The Kaldi speech recognition toolkit." *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[9] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal and S. Khudanpur "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition" in *Proceedings of ICASSP*, Florence, Italy, 2014, pp. 2494-2498.

[10] F. Richardson, D. A. Reynolds and N. Dehak, " A Unified Deep Neural Network for Speaker and Language Recognition," in *Proceedings of Interspeech*, Dresden , Germany, 2015, pp. 1146-1150.

[11] P. A. Torres-Carrasquillo et.al., "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features", *Proceedings of ICSLP*, Denver, Colorado, 2002, pp. 33–36.

[12] F. Richardson, B. Nemsick and D. Reynolds, "Channel Compensation for Speaker Recognition Using MAP Adappted PLDA and Denoising DNNs," in *Proceedings of ODYSS*EY, Bilbao, Spain, 2016, pp. 225-230.

[13] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proceedings of ICASSP*, Florence, Italy, 2014, pp. 4047-4051.

[14] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems,",*Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[15] Y. L. Gwon, W. Campbell, D. Sturim, H. T. Kung, " Language Recognition via Sparse Coding," in *Proceedings of Interspeech*, San Francisco, California, 2016, pp. 2920-2924.