The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 276-286

SUT Submission for NIST 2016 Speaker Recognition Evaluation: Description and Analysis

Hossein Zeinali, Hossein Sameti and Nooshin Maghsoodi

Department of Computer Engineering

Sharif University of Technology, Tehran, Iran

zeinali@ce.sharif.edu, sameti@sharif.edu, nmaghsoodi@ce.sharif.edu

Abstract

In this paper, the most recent Sharif University of Technology (SUT) speaker recognition system developed for NIST 2016 Speaker Recognition Evaluation (SRE) is described. The major challenge in this evaluation is the language mismatch between training and evaluation data. The submission is related to the fixed condition of NIST SRE 2016 and features a full description of the database and the systems which were used during the evaluation. Most of the systems are developed in the context of the i-vector framework. Preconditioning the i-vectors, score normalization methods and the classifier used are discussed. The performance of the systems on the development and evaluation parts of the NIST 2016 SRE16 dataset are reported and analyzed. The best obtained minimum and actual DCF are 0.666 and 0.740, respectively. This is achieved by score fusion of several systems and using different methods for mismatch compensation.

Keywords: Speaker verification, NIST SRE 2016, SUT, i-vector, PLDA

1. Introduction

During the past two decades, National Institute of Standards and Technology (NIST) has organized several speaker recognition evaluations (SRE). The goals of these evaluations are exploring new ideas in speaker recognition and optimizing speaker recognition systems. Like all SREs, in the SRE16 some challenges are followed. One of them is the mismatch between training and evaluation datasets. Due to attention that most of the provided training data is in English while the evaluation data is in Cantonese and Tagalog, efficient methods are required for reducing the effects of this mismatch. The second challenge is short duration enrollment and test utterances. This challenge more happens for test utterances where their duration varies from 10 to 60 seconds. The last challenge is imbalanced multi-session training. In fact, there are two enrollment conditions for SRE16: three segments available for training some speaker models while only one segment for others. The focus of SRE16 is on the telephone speech in Cantonese and Tagalog languages.

In this paper, we provide the description of our system and analyze the results of using different features sets, different Voice Activity Detection (VAD) systems and methods for preconditioning the i-vectors. Our contrastive system is constructed by combining 5 subsystems that each of them is an i-vector based system. The subsystems differ from each other in terms of input features (i.e. MFCC, PLP, SBN or Perseus) or applied VAD method (i.e. FVAD or EVAD). We have developed two sets of these 5 systems with and without labeled data (i.e. Contrastive 1 and 2). The final system is constructed by fusing these two sets. The first version of our system description without any analysis on the evaluation data can be found in [1].

The rest of this paper is organized as follows. We begin with a brief description of those parts of the system which are different from the standard i-vector pipeline. In the next section, the experimental setup for different parts of front-end and back-end, dataset and our subsystems are provided. The performance results are illustrated in Section 4 and finally, in Section 5, we draw conclusions based on the results.

2. System description

In this evaluation, we used i-vector [2] based systems only. Using different features and also different VADs, several systems were trained. All of them used the same Probabilistic Linear Discriminant Analysis (PLDA) [3] back-end. The parts of our system which differ from conventional i-vector framework are explained in the following sub-sections. A schematic block diagram of the system is depicted in Figure 1.

2.1. NAP trained on languages

As mentioned in the introduction, one of the main challenges in this evaluation is the language mismatch between the training and evaluation data. It seems that using a method for reducing the effect

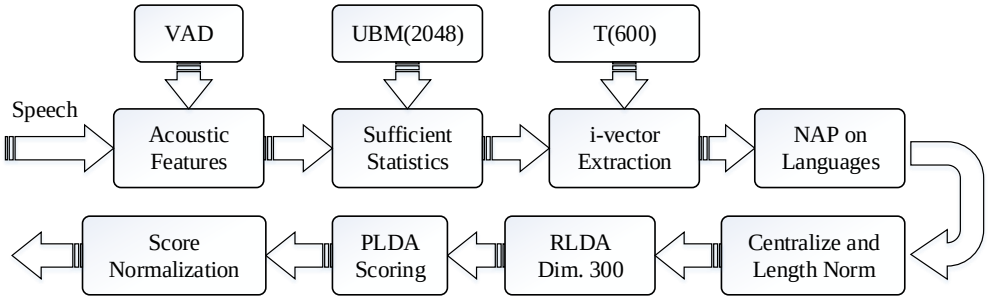

Figure 1: Block diagram of the SUT system with NAP, RLAD, and Score normalization.

of languages may help the performance. Here, in order to reduce the effects of this mismatch, we used Nuisance Attribute Projection (NAP) on top of all i-vectors [2, 4] to project away the language directions. As classes for calculating NAP projection, 20 languages were selected from the primary dataset (see Section 3.1) along with two classes corresponding to the major and minor unlabeled data from the development set. Let $\mathbf{m}_i$ shows the mean of i-vectors for the language i^{th} and $\mathbf{M}_{d\times r}$ shows the matrix of the means (i.e. each column shows mean of one language). If $\mathbf{N} = orth(\mathbf{M})$ be an orthonormal basis for $\mathbf{M}$, then, $\mathbf{A} = \mathbf{I} - \mathbf{N}\mathbf{N}^{'}$ is an square matrix, having $r-1$ eigenvalues equal to zero and $d+1-r$ eigenvalues equal to one. The dimension-reducing projection $\mathbf{P}$ is formed by the eigenvectors associated with the non-zero eigenvalues of $\mathbf{A}$. $\mathbf{P}$ projects away the subspace spanned by all differences between pairs of columns of $\mathbf{M}$.

2.2. Regularized LDA

In addition to NAP projection and prior to training PLDA classifier, i-vectors are centeralized and then length normalized [5]. The centeralizing process has been done by calculating the mean from the primary dataset. Based on our previous works on text-dependent speaker verification [6, 7], Regularized LDA (RLDA) [8] was used instead of using conventional Linear Discriminant Analysis (LDA). In this method, the within and between class covariance matrices are calculated using the following formulas:

$$\mathbf{S}_w = \alpha\mathbf{I} + \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N_s}\sum_{n=1}^{N_s}(\mathbf{w_s}^n - \overline{\mathbf{w}_s})(\mathbf{w_s}^n - \overline{\mathbf{w}_s})^t\,, \quad (1)$$

$$\mathbf{S}_b = \beta\mathbf{I} + \frac{1}{S}\sum_{s=1}^{S}(\overline{\mathbf{w}_s} - \overline{\mathbf{w}})(\overline{\mathbf{w}_s} - \overline{\mathbf{w}})^t\,, \quad (2)$$

where, S is the total number of classes (i.e. speakers in this paper), N_s is the number of training samples in class s^{th}, $\mathbf{w}_s^n$ is the n^{th} sample in class s, and $\overline{\mathbf{w}_s} = \frac{1}{N_s}\sum_{n=1}^{N_s}\mathbf{w}_s^n$ is the mean of class s, $\overline{\mathbf{w}}$ is the mean of total samples, $\mathbf{I}$ is the identity matrix and α and β are two fixed coefficients which have been calculated using the development set.

It is clear that we just add a regularization to each covariance matrix. Alpha and beta parameters are set to 0.001 and 0.01, respectively. Only telephony recordings from the primary data were used for RLDA training. The dimension of i-vectors was reduced to 300 by using RLDA.

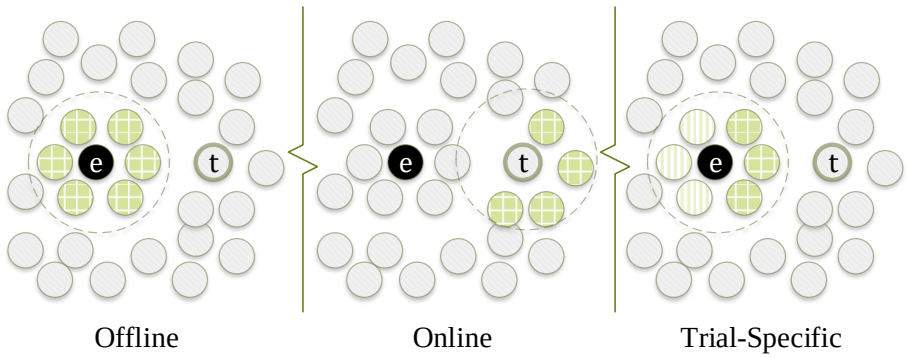

Figure 2: Comparison of the Offline, Online and Trial-Specific methods for imposter set selection. $\mathbf{e}$ is enrollment i-vector and $\mathbf{t}$ is test i-vector. It is clear that the Trial-Specific method selects the i-vectors between two enrollment and test i-vectors as imposter set.

2.3. Score normalization

For score normalization, a specific version of the s-norm method was used. In this method, we used trial specific imposter set selection for t-norm part and offline imposter set selection for z-norm part. During enrollment step, 10000 nearest i-vectors are selected for each model from the primary and unlabeled data. Then, these i-vectors are scored against the model and their mean and standard deviation are used for z-norm. For t-norm, each test i-vector was first scored against these 10000 i-vectors and then, 5000 largest scores were used for calculating mean and standard deviation for t-norm. Since imposter sets depend on both model and test i-vectors, this method is called trial specific. For better intuition, the comparison of this method with the offline and online imposter set selection methods is shown in Figure 2. This figure indicates that in the proposed method, only the i-vectors between the enrollment and the test i-vectors are used as imposter set.

Note that this s-norm method is not symmetric. In the original s-norm method [9], imposter sets for t-norm and z-norm parts are the same and so it is symmetric.

3. Experimental setups

3.1. Dataset

The primary training data is the combination of telephony parts from NIST SRE 2004 - 2008, Fisher English and Switchboard. The unlabeled data from SRE16 development set was used as additional training data. For the final system, we also used labeled data from SRE16 development set. For each subsystem, we used a different subset of these datasets that will be indicated in each section.

3.2. VAD

We did experiments with various VAD methods and based on our findings two of them have been used in this evaluation. Our main VAD is based on a phoneme recognizer system which trained on Fisher

dataset. All frames that recognized as silence or noise were dropped. We will refer to this method by Fisher VAD (FVAD). The secondary VAD is an energy based method that was used in one system. This method is called as Energy VAD (EVAD).

3.3. Features

We used four different feature sets. All acoustic features have 19 coefficients along with Energy that makes 20-dimensional feature vectors. Delta and delta-delta coefficients were also used which makes 60-dimensional feature vectors. These features were extracted using an identical configuration: 25 ms Hamming windowed frames with 15 ms overlap. For each utterance, the features are normalized using short time cepstral mean and variance normalization after dropping the non-speech frames. Three used acoustic features are as follows:

- 19 MFCC + Energy
- 19 PLP + Energy
- Perseus - description of this feature can be found in [10].

Besides the acoustic features, an 80-dimensional DNN based Stacked Bottleneck (SBN) feature was used. This feature was trained using Fisher English dataset. The details about DNN-SBN can be found in [11, 12].

3.4. UBM training

In all systems, a gender-independent diagonal covariance Gaussian Mixture Model (GMM) with 2048 components is used. This model was first trained using about 8000 utterances that were randomly selected from the primary dataset. The MAP adaptation with relevance factor 512 was then used for adapting only means of this model by using unlabeled data from SRE16 development set. Doing in this manner was marginally better than adding unlabeled data to UBM training data.

3.5. i-vector extractor training

In each system, 600-dimensional i-vectors were extracted from original feature sets using a gender-independent i-vector extractor. This component was trained using about 77000 utterances from the primary dataset and unlabeled data from SRE16 development set. It is worth mentioning that for UBM and i-vector extractor training only the telephony data was used.

3.6. Model enrollment

We did some experiments on two common schemes of multi-session enrollment: 1) statistics averaging and 2) i-vectors averaging. The second strategy performed slightly better and so we decided to use it for model enrollment with multiple utterances.

3.7. PLDA

In all systems, we used PLDA as the classifier. The same training data as RLDA is used for PLDA training. The rank of speaker and channel subspaces were set to 200 and 100, respectively.

3.8. Systems

Our final submission is based on 5 i-vector based systems which are different in terms of the input features or VAD:

- 60 dimensional MFCC with EVAD
- 60 dimensional MFCC with FVAD
- 60 dimensional PLP with FVAD
- 60 dimensional Perseus with FVAD
- 140 dimensional MFCC+SBN with FVAD

We did some experiments to find the best strategy for using labeled data from SRE16 development set. When we added this part to RLDA and PLDA training data, we observed a little change in score distributions (i.e. a little shift just on target scores), because the number of speakers in the development set (i.e. 20 speakers) compared to training speakers is very small. As a result, we decided to add this data to the training data of these 5 systems and used them as a complimentary set for the final fusion.

3.9. Final fusion

As mentioned in the introduction, we had two sets of 5 systems. In the first one, we did not add labeled data to the training data, but in the second one, we did. We trained logistic regression for fusion and calibration of each set of systems using BOSARIS toolkit [13]. SRE16 development trials were used for this fusion training. The final submission is the summation of two fused systems (i.e. with and without labeled data).

3.10. System performance

We analyze and compare the systems performance on the SRE16 development and evaluation data using the Equal Error Rate (EER) and the primary cost function. The primary metric in this evaluation is $C_{primary}$, defined as the average cost at two specific points on the DET curve [14]. The detection cost function (DCF) is defined in normalized form as follows:

$$C_{Norm} = P_{Miss|Tar} + \frac{1 - P_{Tar}}{P_{Tar}} \times P_{FalseAlarm|NonTar} \ , \tag{3}$$

where P_{Target} is a priori probability that a trial is a target trial. Actual detection costs will be computed from the trial scores by applying detection thresholds of $log(\beta)$ for the two values of β, with β_1 for

Table 1: Comparison between various methods for MFCC_FVAD. These results were obtained using NIST scoring script in equalized/unequalized modes. S-Norm(Simple) indicates the conventional normalization method compared to the proposed method.

	Development			**Evaluation**		
System name	EER[%]	$min\ C_{Prim}$	$act\ C_{Prim}$	EER[%]	$min\ C_{Prim}$	$act\ C_{Prim}$
LDA+PLDA	21.05 / 21.21	0.884 / 0.897	7.833 / 8.254	15.89 / 16.01	0.971 / 0.978	12.487 / 16.300
RLDA+PLDA	20.82 / 21.15	0.861 / 0.858	6.717 / 6.995	15.62 / 15.81	0.948 / 0.956	10.545 / 13.758
RLDA+PLDA+S-Norm(Simple)	19.29 / 19.52	0.734 / 0.730	0.843 / 0.845	14.01 / 13.62	0.771 / 0.754	0.843 / 0.839
RLDA+PLDA+S-Norm	19.19 / 19.51	**0.718** / 0.713	**0.748** / **0.741**	13.36 / 12.99	**0.731** / **0.706**	0.801 / **0.770**
NAP+RLDA+PLDA+S-Norm	**17.75** / **18.77**	0.719 / **0.699**	0.761 / **0.741**	**12.73** / **12.66**	0.755 / 0.748	**0.771** / 0.783

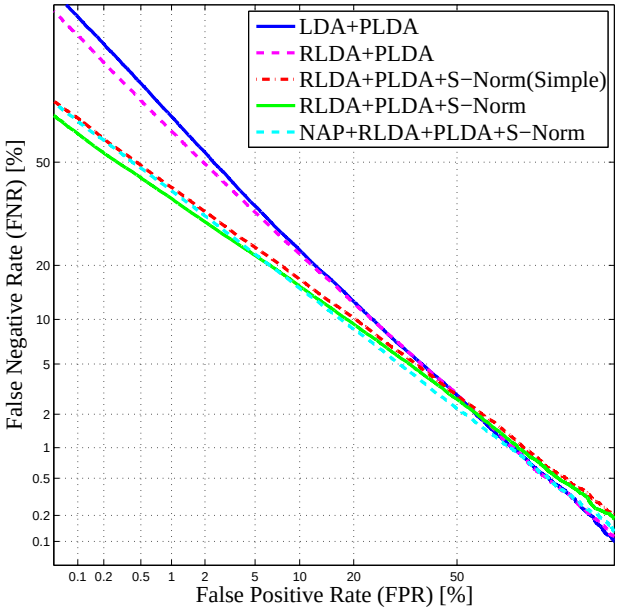

Figure 3: DET plot comparison of different methods for MFCC_FVAD.

$P_{Target_1} = 0.01$ and β_2 for $P_{Target_2} = 0.005$. And finally the primary cost measure for SRE16 is defined as:

$$C_{Primary} = \frac{C_{Norm_{\beta_1}} + C_{Norm_{\beta_2}}}{2}\,. \tag{4}$$

Also, a minimum detection cost will be computed by using the detection thresholds that minimize the detection cost.

4. Results

4.1. Methods comparison

The comparison results for different methods are shown in Table 1. These results were obtained with MFCC features and FVAD. The DET curves of different systems from Table 1 are shown in Figure 3.

By comparing the first and second rows of Table 1, it is clear that RLDA performs better than

conventional LDA in all operating points, especially in actual $C_{primary}$. Similarly, S-Norm improves the performance and also produces calibrated scores. In this evaluation, the effect of the score normalization is higher than in the previous evaluations. The fourth row of Table 1 reports the performance of trial specific imposter set selection algorithm. It is obvious that this method improves the performance in all criteria and its error reduction for evaluation data is more than development data.

The last row of Table 1 reports the effects of NAP method for reducing the effect of language mismatch. Unfortunately, the performance of this method is not consistent with all criteria. It considerably reduces the EER while in most cases it performs worse in the primary cost points. It is clear that the advantages of this method (i.e. EER reduction) is much more than its disadvantage and so we decided to use this method on top of i-vectors.

4.2. Features comparison

Table 2 shows the performance comparison between 5 systems and their fusion for the fixed condition as defined in the SRE16 evaluation plan. It is clear that the PLP system works considerably worse than other acoustic features in terms of EER while it performs about the same in terms of minimum $C_{primary}$. This also happens for SBN features concatenated with MFCCs (i.e. MFCC+SBN). SBN features were trained using Fisher English data and it is proved that the BN features are language dependent and performs the best in the trained language. Although this system performs worst, it helps final fusion in terms of both measures.

One interesting observation from this section is the difference between the minimum and actual primary cost. It is clear that this difference is not so much and this shows well-calibrated scores without any extra calibration method. This is an important advantage of trial specific imposter set selection for score normalization.

The second section of this table reports performance of different systems when labeled data from development set was added to the training data. It is obvious that in this case, considerable improvement is achieved in terms of EER for development set while the improvement of minimum $C_{primary}$ is not so much. The improvement for evaluation data is not as much as the development results because the development set was added to the training data and the systems are over-fitted to this set slightly.

The third section of Table 2 shows fusion results for the first two sections. It is clear that fusion system (i.e. Contrastive 2) performs better than individual systems in terms of both EER and minimum $C_{primary}$. Contrastive 1 performs better than Contrastive 2 on evaluation set while it was over-fitted to development data. This happens because it used in-domain data for training and this reduces the mismatch effects.

The last row of this table shows our final system, which is a simple summation of two systems from the third section. We selected this system as final submission to reduce the possibility of over-fitting effects on evaluation set while Contrastive 1 performs the best.

Table 2: Performance comparison of different systems and their fusion for the SRE16 dataset. These results were obtained using NIST scoring script in equalized/unequalized modes. The first section shows results from single systems without any usage of labeled data. The results in the second section were obtained using the same systems from the first section that used labeled data. Contrastive 1 and 2 are the fusion of systems from second and first sections respectively. The last row shows the final submitted system which is the fusion of two Contrastive systems.

			Development			**Evaluation**		
System name	Lab.	Cal.	EER[%]	$min\,C_{Prim}$	$act\,C_{Prim}$	EER[%]	$min\,C_{Prim}$	$act\,C_{Prim}$
MFCC_EVAD	No	No	**17.54** / **17.96**	0.736 / 0.730	0.782 / 0.773	**12.48** / **12.15**	0.763 / 0.757	**0.767** / **0.769**
MFCC_FVAD	No	No	17.75 / 18.77	**0.719** / **0.699**	**0.761** / **0.741**	12.73 / 12.66	**0.755** / **0.748**	0.771 / 0.783
PLP_FVAD	No	No	19.63 / 20.32	0.773 / 0.781	0.798 / 0.806	13.71 / 14.00	0.782 / 0.789	0.824 / 0.846
Perseus_FVAD	No	No	17.66 / 18.16	0.794 / 0.780	0.811 / 0.799	13.65 / 13.40	0.793 / 0.783	0.809 / 0.813
MFCC+SBN_FVAD	No	No	20.59 / 21.93	0.764 / 0.765	0.803 / 0.806	13.86 / 14.22	0.779 / 0.790	0.790 / 0.804
MFCC_EVAD	Yes	No	15.52 / **15.74**	0.678 / 0.673	0.733 / 0.721	**12.19** / **11.94**	**0.750** / **0.744**	**0.754** / **0.757**
MFCC_FVAD	Yes	No	16.34 / 17.34	**0.666** / **0.654**	**0.708** / **0.692**	12.69 / 12.60	**0.750** / **0.744**	0.769 / 0.782
PLP_FVAD	Yes	No	18.38 / 19.03	0.748 / 0.752	0.756 / 0.763	13.72 / 13.99	0.776 / 0.785	0.823 / 0.844
Perseus_FVAD	Yes	No	**15.37** / 15.99	0.753 / 0.738	0.783 / 0.762	13.41 / 13.16	0.781 / 0.773	0.796 / 0.803
MFCC+SBN_FVAD	Yes	No	19.43 / 20.88	0.741 / 0.741	0.767 / 0.771	13.67 / 14.12	0.774 / 0.785	0.785 / 0.799
Contrastive 1	Yes	Yes	13.26 / 14.04	0.599 / 0.576	0.617 / 0.589	10.41 / 10.15	0.664 / 0.664	0.734 / 0.778
Contrastive 2	No	Yes	15.12 / 15.85	0.642 / 0.616	0.665 / 0.633	10.56 / 10.24	0.671 / 0.670	0.748 / 0.792
Final System	Yes	Yes	14.14 / 14.87	0.620 / 0.598	0.639 / 0.610	10.47 / 10.19	0.666 / 0.666	0.740 / 0.782

Table 3: The final submission results on sub-conditions of evaluation set.

Partition	EER[%]	$min\,C_{Primary}$	$act\,C_{Primary}$
All	10.47 / 10.19	0.666 / 0.666	0.740 / 0.782
Male	10.25 / 09.05	0.620 / 0.573	0.736 / 0.782
Female	10.41 / 10.80	0.708 / 0.730	0.743 / 0.782
Cantonese	05.71 / 06.22	0.514 / 0.531	0.537 / 0.782
Tagalog	15.17 / 14.37	0.807 / 0.803	0.942 / 0.782

4.3. Results on the sub-conditions

Table 3 shows the results of final submission on the sub-conditions of evaluation set. It is obvious that the male and female results are almost the same. The performance of Tagalog is about three times worse than Cantonese in terms of EER. It seems that Tagalog is a more difficult language for speaker verification. Our calibration for male and also for Tagalog is not as good as it was expected.

4.4. Execution time and memory consumption

The reported numbers here were measured using a server with Intel(R) Xeon(R) CPU E5-2640 @ 2.50 GHz and with 64 GB memory.

The most consuming steps in our systems are VAD, feature extraction, and i-vector extraction. For extracting acoustic features, the average execution time of these steps using a single thread is about

13 times faster than real time. This number for MFCC+SBN system is about 2.4 times. The memory consumption for these two system types is 3GB and 5GB respectively.

Although the execution time of enrollment and scoring are negligible with respect to the other steps (i.e. it takes about 1.43 second for one model and 1000 test i-vectors), it is worth noting that our score normalization is slower than conventional s-norm. It needs an extra sorting method before selecting scores for calculating mean and standard deviation.

5. Conclusions

This paper describes SUT system for the fixed condition of NIST SRE16. We used different feature sets and VAD in front-end and made the back-end just based on PLDA. Comparison between features showed that acoustic features perform better than bottleneck features in this evaluation, due to the language mismatch between training and evaluation datasets. NAP is an effective method for reducing the effects of language mismatch but it just helps in EER operating point. Experimental results proved that using RLDA performs better than conventional LDA for preconditioning i-vectors prior to PLDA training.

For score normalization, we have used trial specific imposter set selection method combined with s-norm. This method was the best way for selecting imposter sets. The normalized scores with this method were calibrated well without any additional processing.

Using labeled data from SRE16 development set has a risk of over-fitting. So, our final submission system was the fusion of two fused systems (i.e. with and without using this labeled data). This reduces the possibility of over-fitting. Interestingly, the system that used labeled data performs the best on the evaluation set too.

6. Acknowledgement

The authors would like to thank Brno University of Technology (BUT) for providing setups and systems for running these experiments.

References

[1] Hossein Zeinali, Hossein Sameti, and Nooshin Maghsoodi, "SUT System Description for NIST SRE 2016," *arXiv preprint arXiv:1706.05077*, 2016.

[2] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[3] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[4] William M Campbell, Douglas E Sturim, Douglas A Reynolds, and Alex Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[5] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, 2011, pp. 249–252.

[6] Hossein Zeinali, Hossein Sameti, and Lukas Burget, "HMM-based phrase-independent i-vector extractor for text-dependent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1421–1435, 2017.

[7] Hossein Zeinali, Lukas Burget, Hossein Sameti, Ondrej Glembek, and Oldrich Plchot, "Deep neural networks and hidden Markov models in i-vector-based text-dependent speaker verification," in *Odyssey-The Speaker and Language Recognition Workshop*, 2016.

[8] Jerome H Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[9] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010, p. 14.

[10] Ondřej Glembek, Pavel Matejka, Oldřich Plchot, Jan Pešán, Lukáš Burget, and Petr Schwarz, "Migrating i-vectors between speaker recognition systems using regression neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Pavel Matejka, Ondrej Glembek, Ondrej Novotny, Oldrich Plchot, Frantisek Grezl, Lukas Burget, and Jan Cernocky, "Analysis of DNN approaches to speaker identification," in *ICASSP*, 2016.

[12] Hossein Zeinali, Hossein Sameti, Lukáš Burget, and Jan Černockỳ, "Text-dependent speaker verification based on i-vectors, deep neural networks and hidden Markov models," *Computer Speech & Language*, vol. 46, pp. 53–71, 2017.

[13] Niko Brümmer and Edward de Villiers, "The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing," *Documentation of BOSARIS toolkit*, 2011.

[14] "NIST 2016 speaker recognition evaluation plan," 2016, [Online]. Available at `https://www.nist.gov/sites/default/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf`.