INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speakers In The Wild (SITW): The QUT Speaker Recognition System

H. Ghaemmaghami, M. H. Rahman, I. Himawan, D. Dean, A. Kanagasundaram, S. Sridharan and C. Fookes

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia

{houman.ghaemmaghami, m20.rahman, i.himawan}@qut.edu.au, ddean@ieee.org, {a.kanagasundaram, s.sridharan, c.fookes}@qut.edu.au

Abstract

This paper presents the QUT speaker recognition system, as a competing system in the *Speakers In The Wild* (SITW) speaker recognition challenge. Our proposed system achieved an overall ranking of second place, in the main *core-core* condition evaluations of the SITW challenge. This system uses an i-vector/PLDA approach, with domain adaptation and a deep neural network (DNN) trained to provide feature statistics. The statistics are accumulated by using class posteriors from the DNN, in place of GMM component posteriors in a typical GMM-UBM i-vector/PLDA system. Once the statistics have been collected, the i-vector computation is carried out as in a GMM-UBM based system. We apply domain adaptation to the extracted i-vectors to ensure robustness against dataset variability, PLDA modelling is used to capture speaker and session variability in the i-vector space, and the processed i-vectors are compared using the batch likelihood ratio. The final scores are calibrated to obtain the calibrated likelihood scores, which are then used to carry out speaker recognition and evaluate the performance of the system. Finally, we explore the practical application of our system to the *core-multi* condition recordings of the SITW data and propose a technique for speaker recognition in recordings with multiple speakers.

1. Introduction

Recent advances in speaker recognition technology, specifically text-independent speaker verification, have brought about significant gains in system accuracy. The joint factor analysis (JFA) speaker modelling approach proposed by Kenny [1], has evolved into a powerful tool for speaker verification. This is because the JFA approach allows for modelling of inter-speaker variability and compensation for channel/session variability in the context of high-dimensional Gaussian mixture model (GMM) supervectors. This technique advanced to a new front-end factor analysis technique, termed i-vector (for intermediate-size vector) extraction, proposed by Dehak *et al.* [2]. In this technique, rather than taking the JFA approach of modelling a speaker and channel variability spaces, a low-dimensional total-variability space that models both speaker and channel variability is trained. The i-vector approach proposed in [2], has the advantage of scoring using a cosine similarity scoring (CSS) kernel directly to perform verification, making the scoring process faster and less complex than other speaker verification methods, including JFA or support vector machines (SVM).

The use of i-vectors for speaker modelling has been established as the state-of-the-art approach to representing speech segments produced by a speaker identity, however i-vectors are susceptible to unwanted variations due to mismatch of linguistic content and recording channel information between segments of speech spoken by the same speaker identity. To overcome this, a range of techniques, such as within-class covariance normalisation (WCCN), linear discriminant analysis (LDA) and nuisance attribute projection (NAP) were proposed and shown to be effective for this purpose [3]. Kenny *et al.* [4] then introduced the use of the PLDA approach for modelling channel variability within the i-vector space. The PLDA technique was originally proposed by Price *et al.* [5] for face recognition, and later it was adapted for modelling the i-vector distributions for speaker verification [6]. Two PLDA approaches, Gaussian PLDA (GPLDA) and heavy-tailed PLDA (HTPLDA) were introduced [7, 6]. Garcia-Romero *et al.* [6], demonstrated that the heavy-tailed behaviour of i-vector features can be converted into Gaussian behaviour by using a length-normalized approach, and thus length-normalized GPLDA was demonstrated to achieve similar performance to the HTPLDA technique.

Until recently the use of a GMM universal background models (UBM), trained on large amounts of unlabelled data for capturing phonetic variations of speech in an unsupervised manner, was the state-of-the-art approach and an essential part of speaker recognition technology. The successful application of deep neural network (DNN) acoustic models to the task of automatic speech recognition (ASR) [8] and the significant gains in accuracy that resulted from the use of a DNN for this task, has brought about proposals for the use of such a DNN in the task of speaker recognition [9, 10]. In order to apply such an ASR DNN to the task of speaker recognition, the GMM-UBM is replaced by a DNN to collect sufficient statistics for i-vector speaker modelling [10]. The input to the DNN is frequency-domain features, such as the mel-frequency cepstral coefficient (MFCC) features, while the output of the DNN provides soft alignments for phonetic content, in the form of tied triphone states, referred to as senones [9]. This technique has resulted in significant improvements over GMM-UBM i-vector systems [9]. It is hypothesised that this is due to the ability of the DNN to model phonetic content/variations directly, as opposed to the unsupervised expectation-maximization (EM) approach in the GMM-UBM training process for capturing some acoustic patterns in the data, which would highly depend on the training data. This however is achieved at the cost of higher computational complexity and can put significant strain on resources.

Another factor that can significantly impact the performance of speaker verification systems is the mismatch between audio domains of the training and test data. The performance variation due to cross-domain speaker verification, was first addressed at the Summer Workshop at Johns Hopkins University (JHU) held in 2013 [11]. Results presented in that

http://dx.doi.org/10.21437/Interspeech.2016-945

workshop clearly highlighted the performance gap between in-domain and out-domain development for speaker verification, which must be addressed. This task was deemed the *Domain Adaptation Challenge* (DAC) [11]. We use a dataset invariant covariance normalization (DICN) approach to compensate the mismatch between in-domain and out-domain datasets in the i-vector space. Instead of capturing the mismatch directly between out-domain and in-domain data [12], we capture the mismatch as compared to a global mean i-vector [13]. This is detailed further in the paper.

We proposed and developed a state-of-the-art DNN i-vector speaker verification system, with inter-dataset variability PLDA [13]. We submitted our system's evaluation results to the *Speakers In The Wild* (SITW) 2016 speaker recognition challenge [14]. Our proposed system achieved a ranking of second place, out of all participating teams, in the main *core-core* condition evaluations across the SITW evaluation data. In this condition, a segment of audio with speech from a single speaker (but including potential non-speech and noise) is compared to another segment of speech from a claimant (also possibly including non-speech and noise). This paper provides a detailed system description of our submitted system, providing details of our DNN, i-vector extraction process, inter-dataset variability compensation, i-vector scoring and score calibration. We detail our training and development data, report our results obtained across the *core-core* evaluation conditions of the SITW test data and discuss the limitations of applying a DNN i-vector system to the *core-multi* condition of the SITW evaluation data [14], proposing a possible strategy.

2. Data used for training and development

In order to train a DNN for extracting feature statistics, we use 300 hours of data from the Fisher corpus [15].

We use the NIST SRE 2004, 2005, 2006 and 2008 datasets, as well as Switchboard phase II and III, and Switchboard Cellular Part 2 corpora for training the total variability matrix for i-vector extraction.

For inter-dataset variability compensation in the i-vector space [12, 13], we compute the mean i-vector from 651 speakers in the SITW development set.

For PLDA training, we pooled together telephone conversation data from the NIST SRE 2004, 2005 and 2006 datasets, together with Switchboard I and II, as well as Switchboard phase I, II and III corpora. We used data from a total number of 4800 speakers, across 32095 sessions.

In order to carry out score calibration, after obtaining the full set of scores over all SITW *core-core* development data, the BOSARIS toolkit [16] was used to calibrate the evaluation scores into true log-likelihood-ratios using the non-parametric PAV (pool adjacent violators) approach [16], based on the scores returned on the labelled SITW *dev* dataset.

3. DNN for collecting feature statistics

The recent use of ASR DNNs in place of a GMM-UBM for collecting sufficient statistics, for i-vector training in the task of speaker recognition, has brought about significant improvements [8, 9]. Waibel *et al.* first proposed the use of time-delay neural networks (TDNN) for phoneme recognition [17]. This motivated the work by Snyder *et al.* [10], which proposed the use of such a DNN for obtaining sufficient statistics for i-vector training. TDNNs are capable of dealing with long temporal contexts and can hence better model the long-term changes of acoustic events in speech [17, 10].

We use the Kaldi toolkit [18] to develop our DNN for collecting feature statistics. We follow the DNN recipe proposed for speaker recognition by Snyder *et al.* [10], and train a multi-splice TDNN with six hidden layers and a splicing configuration. The hidden layers use a *p*-norm activation function (where *p*=2). The input layer takes 40 dimensional MFCC features with 5-frame temporal context and cepstral mean subtraction (CMS) performed over a window of 6 seconds. The features are extracted every 25ms with a 10ms window shift. Each hidden layer has 350 nodes, the output dimension is 3500 and a softmax output layer computes posteriors for 5,346 senone targets [10, 9]. The forced alignment between the state-level transcripts and the corresponding speech signals by the GMM/HMM triphone system is used to generate labels for DNN training.

As the non-speech segments of SITW data can negatively impact system accuracy, we use the energy-based voice activity detection (VAD) provided in the Kaldi toolkit [18] to carry out VAD prior to extracting the sufficienct statistics using our trained DNN. This approach uses an energy threshold across the zero coefficient of extracted MFCC features to carry out VAD. We use a threshold value of 5.5.

4. i-vector based speaker verification

Rather than decomposing a GMM mean supervector into separate channel and speaker components, as in the JFA approach, Dehak *et al.* [2] proposed the representation of a GMM mean supervector in a single space that capture all variabilities. This was motivated by the discovery that the channel space of JFA still contains information that can be used to aid the task of speaker verification.

In the i-vector approach, both the speaker and channel dependent GMM supervectors are represented by a single i-vector and its projection based on the total variability space,

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{m}$ is the speaker and session independent UBM mean supervector, $\mathbf{T}$ is a low rank matrix referred to as the total variability matrix. $\mathbf{w}$ is the total variability factor (or i-vector) which is normally distributed. A detail procedure of total-variability subspace ($\mathbf{T}$) training and i-vector extraction is described in [2, 19].

We employ the Kaldi toolkit [18] to compute and extract 600-dimensional i-vectors using the feature statistics provided by our trained DNN (detailed in Section 3). We use 20-dimensional MFCC features for extracting i-vectors, with the energy-based VAD (as described in Section 3) applied as the front-end processing stage.

5. Inter-dataset variability compensation

As we use telephone conversation data for training, it is necessary to minimise the impact of dataset variability on the accuracy of our system. To capture potential dataset variability, we apply inter-dataset variability compensation in the i-vector space. In this approach dataset variability is captured using the outer product of the difference between the training i-vectors and the mean of the i-vectors extracted from speakers in the SITW-dev dataset [12, 13]. The training dataset is then projected into a new, dataset variability compensated, subspace.

$$\mathbf{\Sigma}_{DVC} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}_n - \bar{\mathbf{w}})(\mathbf{w}_n - \bar{\mathbf{w}})', \tag{2}$$

where N is the total number of development i-vectors and $\bar{\mathbf{w}}$ is the mean of combined set of NIST and SITW i-vectors, which can be calculated as follows,

$$\bar{\mathbf{w}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{w}_i. \tag{3}$$

The training dataset is then projected into a new, dataset variability compensated, subspace.

The matrix $\mathbf{A}$ is used to first scale the subspace, where $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}_{DVC}^{-1}$. The dataset variability compensated development i-vectors are then extracted as follows,

$$\hat{\mathbf{w}} = \mathbf{A}^T\mathbf{w}_{dev}. \tag{4}$$

6. Linear discriminant analysis (LDA)

LDA is a channel compensation method [2, 20], which attempts to find the orthogonal directions in the feature space to minimize the intra-class variance caused by channel and maximize the variance between speakers through the eigenvalue decomposition of,

$$\boldsymbol{\Sigma}_b\mathbf{v} = \tau\boldsymbol{\Sigma}_w\mathbf{v}, \tag{5}$$

where τ is the eigenvalues, $\mathbf{v}$ is the eigenvector, $\boldsymbol{\Sigma}_w$ is within class matrix and $\boldsymbol{\Sigma}_b$ is between class matrix.

The between-and within-class covariance matrices are determined as follows,

$$\boldsymbol{\Sigma}_b = \sum_{s=1}^{S} n_s(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})', \tag{6}$$

$$\boldsymbol{\Sigma}_w = \sum_{s=1}^{S}\sum_{1=1}^{n_s}(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)', \tag{7}$$

where S is the total number of out-domain speakers, n_s is the number of sessions of speaker s. $\bar{\mathbf{w}}_s$ is the mean i-vector for each speaker and $\bar{\mathbf{w}}$ is the mean of all speakers which are defined by,

$$\bar{\mathbf{w}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{w}_i^s, \tag{8}$$

$$\bar{\mathbf{w}} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\mathbf{w}_i^s, \tag{9}$$

where N is the total number of sessions.

In the low dimensional space resulting from the linear transformation $\mathbf{G}$, the within class and between class matrices become $\boldsymbol{\Sigma}_w = \mathbf{G}^T\boldsymbol{\Sigma}_b\mathbf{G}$. An optimal transformation matrix $\mathbf{G}$ is then trained to maximise trace ($\boldsymbol{\Sigma}_b$) and minimise ($\boldsymbol{\Sigma}_w$),

$$\max_{G}\{\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_b\}. \tag{10}$$

The LDA projected i-vectors are calculated as follows,

$$\mathbf{w}_{LDA} = \mathbf{G}^T\mathbf{w}. \tag{11}$$

7. Length-normalized GPLDA system

Rather than compensating for channel variability in the i-vector space using the subspace transformation approach, a more generative approach heavy-tailed (HTPLDA) is introduced by Kenny [7] to model session and channel variability in the i-vector space. But Garcia-Romero *et al.* [21] have introduced a length-normalized approach to convert the heavy-tailed behaviour to Gaussian behaviour, which is in performance comparable to, but computationally more efficient than, heavy-tailed (HTPLDA).

In this approach the non-Gaussian behaviour of i-vector features are converted into Gaussian i-vector feature behaviour. This technique consists of two steps: (1) linear whitening and (2) length normalization. A linear-whitened i-vector $\mathbf{w}_{LDA-wht}$ can be estimated as follows,

$$\mathbf{w}_{LDA-wht} = \mathbf{d}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{w}_{LDA}, \tag{12}$$

where $\mathbf{U}$ is an orthonormal matrix containing the eigenvectors and $\mathbf{d}$ is a diagonal matrix containing the corresponding eigenvalues.

A length-normalized i-vector $\mathbf{w}^{norm}$ can be found as follows,

$$\mathbf{w}_{LDA}^{norm} = \frac{\mathbf{w}_{LDA-wht}}{\|\mathbf{w}_{LDA-wht}\|}. \tag{13}$$

A speaker and channel dependent length-normalized i-vector can be defined as,

$$\mathbf{w}_{LDA-r}^{norm} = \mathbf{w}_{LDA}^{norm} + \mathbf{U}_1\mathbf{x}_1 + \epsilon_r, \tag{14}$$

where for given speaker recordings $r = 1, 2, ...R$; $\mathbf{w}_{LDA}^{norm} + \mathbf{U}_1\mathbf{x}_1$ is the speaker specific part and ϵ_r is the channel specific component; The covariance matrix of the speaker component is $\mathbf{U}_1\mathbf{U}_1^T$ and the covariance matrix of the channel component is $\boldsymbol{\Lambda}^{-1}$. Training of the eigenvoice matrix $\mathbf{U}_1$ is the same as learning the eigenvoice matrix $\mathbf{V}$ in the JFA modelling approach [2, 1].

GPLDA scoring is calculated using the batch likelihood ratio [7]. Given target i-vectors $\mathbf{w}_{target}$ and test i-vectors $\mathbf{w}_{test}$, batch likelihood ratio can be calculated as follows,

$$\ln\frac{P(\mathbf{w}_{target}, \mathbf{w}_{test} \mid H_1)}{P(\mathbf{w}_{target} \mid H_0)P(\mathbf{w}_{test} \mid H_0)}, \tag{15}$$

H_1: The speakers are same, H_0: The speaker are different.

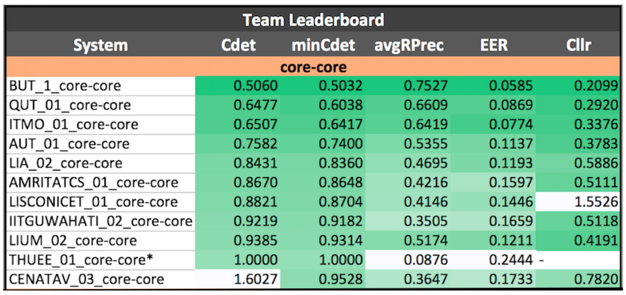
8. Results and discussions

We first evaluated our approach across all data from the development set of the SITW *core-core* condition [14], obtaining batch likelihood scores for each development trial. The obtained scores were then assessed against the development keys and used for score calibration of the batch likelihood scores obtained in the *core-core* condition evaluations. This was done using the approach described in Section 8.

In this section we first present our system's performance across the SITW evaluation data, as assessed by the organisers of the SITW 2016 challenge [14]. We then discuss our efforts in addressing the *core-multi* condition evaluations and discuss the pitfalls of applying a DNN i-vector/PLDA approach to such computationally demanding tasks.

8.1. SITW *core-core* evaluations

Figure 1 displays the team leaderboard for the main *core-core* evaluations across the SITW dataset. Our proposed system, the QUT speaker recognition system, is displayed as the *QUT_01_core-core* system. It can be seen that our proposed approach achieved an overall ranking of second place. We submitted two identical systems to the challenge, with the difference



Team Leaderboard					
System	Cdet	minCdet	avgRPrec	EER	Cllr
core-core					
BUT_1_core-core	0.5060	0.5032	0.7527	0.0585	0.2099
QUT_01_core-core	0.6477	0.6038	0.6609	0.0869	0.2920
ITMO_01_core-core	0.6507	0.6417	0.6419	0.0774	0.3376
AUT_01_core-core	0.7582	0.7400	0.5355	0.1137	0.3783
LIA_02_core-core	0.8431	0.8360	0.4695	0.1193	0.5886
AMRITATCS_01_core-core	0.8670	0.8648	0.4216	0.1597	0.5111
LISCONICET_01_core-core	0.8821	0.8704	0.4146	0.1446	1.5526
IITGUWAHATI_02_core-core	0.9219	0.9182	0.3505	0.1659	0.5118
LIUM_02_core-core	0.9385	0.9314	0.5174	0.1211	0.4191
THUEE_01_core-core*	1.0000	1.0000	0.0876	0.2444	-
CENATAV_03_core-core	1.6027	0.9528	0.3647	0.1733	0.7820

Figure 1: SITW 2016 Speaker Recognition Challenge leaderboard, indicating the second place ranking achieved by the QUT system in the main *core-core* track. The leaderboard contains the best submission, from each team per condition, ranked by the primary metric Cdet (http://www.speech.sri.com/projects/sitw/).

in the two systems being in the application of inter-dataset variability (IDV) compensation. Our system that is ranked in the leaderboard is our best system, which includes IDV compensation. We believe one of the key components of our system that provides significant accuracy is our proposed IDV PLDA technique [13], detailed in Section 5.

8.2. Exploring the SITW *core-multi* evaluations

The *core-multi* evaluation track of the SITW challenge requires the enrolment of a recording containing speech from a single speaker identity, which is then needed to be compared to an unlabelled recording containing speech from multiple speakers. The goal of the task is then to identify if the enrolled speaker also appears in the multiple-speaker, trial recording [14].

We attempted to fully participate in this evaluation track, however due to the computationally expensive nature of the DNN approach, limitation of resources and a lack of time, we were limited to exploring this track. For this reason, we first begin by proposing a practical approach to speaker verification under the *core-multi* evaluation conditions and report on the results that we were able to complete within the limited timeframe. We believe our approach has the potential to provide competitive performance and thus see merit in providing a brief description and discussion in this section.

The comparison of an enrolment recording to a trial recording, for speaker verification, is straightforward across the *core-core* data. This becomes more difficult when dealing with the *core-multi* condition data. One solution is to segment the trial recording into speaker-homogeneous segments. These segments can then be compared and scored against the enrolled speaker to make a decision regarding an identity match.

We draw from our speaker diarization research in [22] and propose the use of an ergodic hidden Markov model (HMM) for segmenting the trial recording into speaker-homogeneous segments. We first mark arbitrary speaker-change points in the trial recording, at 10 second intervals. We then model each of these segments as a state within an ergodic HMM. Each state/segment is modelled using a 32-component GMM [22]. We then carry out three iterations of Viterbi segmentation, with a minimum duration of 2.5 seconds, to refine the segment boundaries. It must be noted that as we use an ergodic HMM, merging of states/segments is also valid for non-adjacent segments, across the entire trial recording. Figure 2 displays the architecture of our HMM segmentation system. To take further advantage of the ergodic architecture of this system, we include the enrolment recording as an additional state in the segmentation process. As we know that the enrolment data only belongs to a single speaker, we can use this constraint to leverage the information in the enrolment recording for (ideally) improving the segmentation of the person of interest (POI) in the trial audio. After segmentation is carried out, each unique and ideally speaker-homogeneous segment in the trial recording is modelled using a GMM-UBM i-vector/PLDA approach and batch likelihood scoring is carried out to score each of these segments against the enrolment recording. It would be a sound assumption that (after voice activity detection) a match between a trial segment and the enrolment data would achieve the best batch likelihood score, out of all possible scores between the enrolment recording and trial recording segments. We thus select the best possible batch likelihood score from this set, as the likelihood of the presence of a POI in the trial recording.

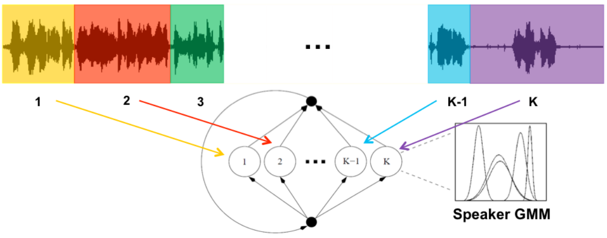


Figure 2: Ergodic HMM segmentation approach for segmenting the trial recording, where each state is a GMM and state 1 represents the enrolment data that is used to improve segmentation of the POI in the trial recording.

We were able to complete a total of 1,996,823 trials of the *core-multi* evaluations, for which we achieved an overall EER of 11.1%. As score calibration would require the processing of the entire development set in the *core-multi* condition trials, we were unable to achieve a complete system to submit for ranking in this condition category *core-multi*. It must be noted that we did not apply inter-dataset variability compensation in our *core-multi*, which may have improved our results further.

9. Conclusions

We proposed the QUT speaker recognition system and were successful in achieving a second place ranking in the main *core-core* condition of the SITW 2016 evaluations. This system is based on a DNN i-vector/PLDA system, with inter-dataset variability compensation used to improve cross-domain evaluations. We also explored the *core-multi* evaluation track of the SITW challenge data. We proposed the use of an ergodic HMM segmentation approach, that takes advantage of the enrolment data to segment the trial recording into speaker-homogenous segments that can be assessed against the enrolment data to make a verification decision. We aim to further explore and analyse the various tracks of the SITW data in our future work.

10. Acknowledgements

This research was supported by an Australian Research Council (ARC) Linkage Grant No: LP130100110.

11. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 2005, pp. 637–640.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. 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] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine Similarity Scoring without Score Normalization Techniques," in *Odyssey The Speaker and Language Recognition*, Brno, Czech Republic, 2010.

[4] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 7649–7653.

[5] J. R. Price and T. F. Gee, "Face recognition using direct, weighted linear discriminant analysis and modular subspaces," *Pattern Recognition*, vol. 38, no. 2, pp. 209 – 219, 2005.

[6] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *ICASSP*, 2012.

[7] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 2010, p. 14.

[8] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=171498

[9] Y. Lei, L. Ferrer, M. McLaren *et al.*, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[10] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 92–97.

[11] (2013) Domain adaptation challenge 2013, johns hopkins university,. [Online]. Available: http://www.clsp.jhu.edu/workshops/archive/ws13-summerworkshop/groups/spk-13/.

[12] A. Kanagasundaram, D. Dean, and S. Sridharan, "Improving outdomain plda speaker verification using unsupervised inter-dataset variability compensation approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4654–4658.

[13] M. H. Rahman, A. Kanagasundaram, D. Dean, and S. Sridharan, "Dataset-invariant covariance normalization for out-domain plda speaker verification," *Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015*, pp. 1017–1021, 2015.

[14] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *submitted to Interspeech*, March 2016.

[15] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text."

[16] N. Brummer and E. de Villiers, "The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing," AGNITIO Research, South Africa, Tech. Rep., nov 2011.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 3, pp. 328–339, 1989.

[18] D. Povey, "The Kaldi speech recognition toolkit." in *in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[19] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[20] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey Speaker and Language Recogntion Workshop*, 2010, pp. 28–33.

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

[22] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker attribution of australian broadcast news data," *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM): CEUR Workshop Proceedings, Volume 1012*, pp. 72–77, 2013.