Effect of long-term ageing on i-vector speaker verification

Finnian Kelly[1], *Rahim Saeidi*[2], *Naomi Harte*[1], *David van Leeuwen*[3]

[1] Department of Electronic & Electrical Engineering, Trinity College Dublin, Ireland
[2] Speech and Image Processing Unit, School of Computing, University of Eastern Finland
[3]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
kellyfp@tcd.ie, rahim.saeidi@uef.fi, nharte@tcd.ie, d.vanleeuwen@let.ru.nl

Abstract

Assessing the impact of ageing on biometric systems is an important challenge. In this paper, an i-vector speaker verification framework is used to evaluate the impact of long-term ageing on state-of-the-art speaker verification. Using the Trinity College Dublin Speaker Ageing (TCDSA) database, it is observed that the performance of the i-vector system, in terms of both discrimination and calibration, degrades progressively as the absolute age difference between training and testing samples increases. In the case of male speakers, the equal error rate (EER) increases from 4.61% at an ageing difference of 0–1 years to 32.74% at an age difference of 51–60 years. The performance of a Gaussian Mixture Model - Universal Background Model (GMM-UBM) system is presented for comparison. It is shown that while the i-vector system outperforms the GMM-UBM system, as absolute age difference increases, the performance of both degrades at a similar rate. It is concluded that long-term ageing variability is distinct from everyday intersession variability, and therefore must be dealt with via dedicated compensation strategies.

Index Terms: speaker verification, ageing, i-vector, GMM-UBM

1. Introduction

The process of ageing leads to gradual changes in the voice throughout adulthood. The impact of ageing on the properties of the voice has been well documented [1, 2, 3, 4, 5]. As biometric systems grow in their reach and scale, establishing the impact of ageing is becoming increasingly important [6, 7]. However, the effect of ageing on speaker verification has not received significant research attention.

In our previous papers [8, 9, 10], and in studies by Rhodes [11, 12], long-term ageing is demonstrated to impact negatively on speaker verification and forensic automatic speaker recognition, leading to errors in classification and weight-of-evidence computation for most speakers within a time-span of 10 years. Prompted by this, proposals for ageing compensation were presented in [8, 9].

In our previous studies, a Gaussian Mixture Model - Universal Background Model (GMM-UBM) approach [13] has been used exclusively. There have been significant developments in speaker verification research in recent years, with progress driven by the regular NIST speaker recognition evaluations (SREs) [14, 15]. The current wave of systems, the majority of which build upon the GMM-UBM framework, incorporate various techniques to improve performance in the presence of inter-session variability. As a result, they have reached a level of performance that significantly outperforms the 'classic' GMM-UBM approach in challenging conditions; for example, see the comparison of GMM-UBM and Joint Factor Analysis (JFA) presented in Kinnunen's review paper [16]. A current research trend [17] is the use of an i-vector framework [18] with PLDA (probabilistic linear discriminant analysis) [19].

Optimisation of speaker verification performance is not the focus of this paper; what is of interest however, is to evaluate how these recent developments, aimed at dealing with intersession variability, behave when faced with ageing variability. Thus, we employ an i-vector system with PLDA modelling, developed at Radboud University Nijmegen (RUN) for the NIST SRE 2012 evaluation [20, 15]. A speaker verification experiment using the Trinity College Dublin Speaker Ageing (TCDSA) database is designed to observe the extent to which this 'state-of-the-art' system is affected by long-term ageing variability. The performance of a GMM-UBM system on the same evaluation is presented for comparison.

2. Speaker Ageing Data

The Trinity College Dublin Speaker Ageing (TCDSA) Database is a longitudinal speech corpus containing recordings of 26 speakers (15 males and 11 females) with an age difference range of 28–58 years per speaker. The most recent version of TCDSA [9] is used for the experiments in this paper. The data was obtained from a variety of sources; however, all recordings are professional radio or television broadcasts. In previous papers, e.g., [8, 9], TCDSA recordings were screened subjectively and objectively to limit non-ageing-related variability, a process detailed fully in [8]. Here however, the *objective* screening measure was not applied, maximising the quantity of data under test.

Accompanying the main database is the TCDSA-UBM database, containing 30 seconds of speech from each of 120 speakers, balanced across gender, age and accent, and containing recordings comparable to those of the main database. In addition, male and female speakers from an expanded version of the TCDSA-FD (Forensic Development) database [10] were used for normalization purposes. All data used in this paper is freely available for academic research[1].

3. i-vector system description

An i-vector system developed for NIST SRE 2012 at Radboud University Nijmegen (RUN) [20, 17] was used in an 'off the shelf' manner for the experiments in this paper. It consists of a standard i-vector [18] configuration with PLDA modelling [21].

[1] To access the data and associated documentation, see: `http://www.sigmedia.tv/Research/SpeakerVerification`

All speech data used was at 8 kHz (downsampling was applied where necessary). The speech signal was enhanced by applying a Wiener filtering based module to the magnitude spectrum of the frames, with the noise spectrum estimated using an improved minima controlled recursive averaging (IMCRA) approach [22]. The front-end consisted of 19-dimensional MFCC (plus log energy) extraction over 20 ms windows every 10 ms. Delta and acceleration coefficients, computed over 9 consecutive frames, were then appended. Speech activity detection was applied according to a Gaussian modelling of the frame energy [23]. Lastly, feature warping [24] was applied.

Gender-dependent UBMs of 2048 components were trained using segments from the following datasets: NIST SRE 2004-2006, Switchboard cellular phase 1 and 2 and Fisher English [17]. An i-vector extractor matrix T of rank 400 was estimated using the same utterances used to train the UBM. Baum-Welch statistics of 0th, 1st and 2nd order are computed using the UBM, and along with the T matrix, were used to extract i-vectors for the relevant utterances.

To reduce intra-speaker variability and enhance inter-speaker variability, LDA (linear discriminant analysis) was applied to the i-vectors, reducing their dimensionality to 200. Finally, the i-vectors are centred, whitened [25] and length-normalized [26]. The speaker and session dependent i-vector distribution was modelled with PLDA [19]. PLDA development data was drawn from NIST SRE 2006-2012 according to the I4U development lists [17]. The score for each trial is the log-likelihood ratio of the pair of i-vectors originating from the same speaker versus different speakers. Score calibration was not applied.

4. GMM-UBM system description

A GMM-UBM system configuration, consistent with our previous work on ageing speaker verification [9], was used in this paper. All speech was preprocessed by downsampling to 8 kHz, applying pre-emphasis and removing silences with an energy-based speech activity detector. Feature extraction consisted of 12-dimensional MFCC extraction over 20 ms windows every 10 ms. Delta coefficients, computed over 5 consecutive frames were appended. RASTA filtering, and mean and variance normalisation were then applied to the feature vector [16]. A gender-independent UBM of 512 components was trained with the TCDSA-UBM database. Speaker GMMs were trained with mean-only adaptation. A relevance factor of 16 was used, and all UBM components were considered in adaptation and scoring.

A log-likelihood ratio (LLR) score was computed for each trial from the likelihoods of the test sample given both the speaker GMM and the UBM. Z-norm [16] was applied to each score. The Z-norm statistics for each speaker GMM were estimated given a set of 25 speakers (of the same gender) from the TCDSA-FD database.

5. Experimental evaluation

A ‘forwards’ and ‘backwards’ verification approach has generally been adopted in our previous papers, e.g., [27, 8, 9], where a speaker’s youngest and oldest recordings are used in training, and the remainder of their recordings are reserved for testing. In this paper, the experimental protocol is expanded by using speaker recordings at all ages for both training and testing, maximising the number of trials. There were a few constraints to this all-vs-all protocol: given the widely variable recording durations within the database, all training and testing samples were restricted to 30 seconds, and a maximum of five samples from a recording were used for testing. No same-session trials were considered at the results analysis stage.

5.1. Experimental results

A feature of the TCDSA database is that the number of recordings per speaker, and their distribution across age, is variable. This makes performance evaluation challenging. Figure 1 illustrates the distribution of trials across speaker and age difference range with our evaluation protocol; it is evident that a subset of the speakers dominate the trials (note the log scale on the y-axis), and that the number of trials generally decreases with age difference.

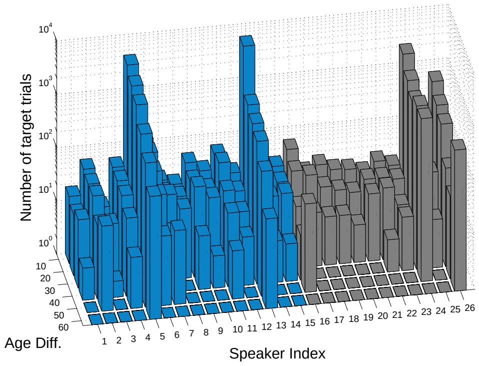

Figure 1: The number of target trials for each speaker at increasing age difference ranges, given the evaluation protocol adopted in this paper. ‘Age Diff.’ denotes the upper limit of a ten year absolute age difference range, e.g., Age Diff. = 20 indicates an age difference range of ±11–20 years between training and testing samples. The indices of male and female speakers are 1–15 and 16–26 respectively. The distribution of non-target trials is similar

In assessing the performance of the speaker verification systems therefore, it is necessary to account for the trial ‘imbalances’ in Figure 1. In [28], van Leeuwen presents a method for calculating the equal error rate (EER) in a way that balances the contributions of different conditions in an evaluation. The false acceptance rate (FAR) and false rejection rate (FRR) are determined for each condition individually and combined. This has the effect of weighting each trial by the inverse of the number of trials of that condition. A similar trial weighting scheme is applied in [29]. In the present case, speakers or age difference ranges can be considered as different ‘conditions’.

A comparison of unweighted EERs with EERs weighted by speaker and by age difference range (the six ranges in Figure 1), is provided in Table 1. The effect of weighting is apparent in all cases, shifting the EER by 1–2% in the speaker weighted case and by 1–8% in the age weighted case. Comparing the speaker weighted EERs, the i-vector system outperforms the GMM-UBM system, as anticipated, with an absolute EER difference between the systems of approximately 7% and 2% for males and females respectively. For both systems, female EER is higher than male EER. Corresponding DET curves for the speaker weighted case are provided in Figure 2.

To evaluate the effect of increasing age difference between training and testing samples, the speaker weighted EERs were evaluated for each system at seven absolute age difference

		pooled	s.weighted	a.weighted
male EER%	i-vec.	13.12	11.52	21.17
	GMM	16.73	18.21	21.68
female EER%	i-vec.	14.82	16.69	16.77
	GMM	19.35	18.59	20.83

Table 1: EERs for unweighted trials ('pooled') and for trials weighted per speaker ('s.weighted'), and per age difference range ('a.weighted'). Trials at all absolute age differences (±0–60 years) are considered.

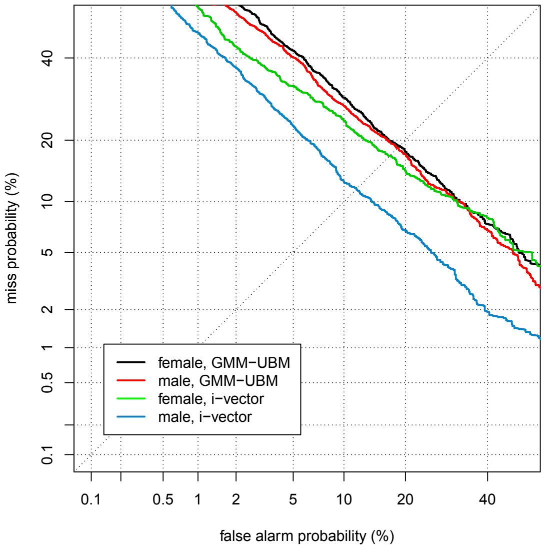

Figure 2: DET curves, weighted per-speaker, for the i-vector and GMM-UBM systems given trials at all absolute age differences (±0–60 years).

ranges, and are shown in Table 2. The first range, 0–1, is based on all trials with an absolute age difference of less than or equal to one year (excluding same-session trials). The effect of ageing over a one year period is assumed to be minimal, and thus the resulting EERs can be considered as ageing-independent baselines. The i-vector system clearly outperforms the GMM-UBM system in the male case. In the female case however, the GMM-UBM EER is slightly lower. At the 0–1 age range, there are zero trials for several female speakers, and a low number of target trials overall (441). This likely contributes to the discrepancy in i-vector performance at this range.

The speaker weighted EER values in Table 2 generally follow an increasing trend as the age ranges increase. In the male case, the EER for the i-vector system at the range 0–10 is lower than the GMM-UBM system, as would be expected based on Figure 2. For subsequent age ranges, the EER increases for both systems at approximately the same rate. At the final range 51–60 however, the GMM-UBM EER drops below that of the i-vector system. In the female case, there is little difference between the EERs of both systems at each age range. For both, there is an increase up to 31–40 range, followed by a decrease to the final age range. For the final four age ranges, the GMM-UBM EER is lower than that of the i-vector system. The 'spike' in the EER at 31–40 is likely attributable to the particular distribution of speakers at this age range.

In addition to the EER, a detection cost metric [14] was defined as:

$$C_{\mathrm{det}}(\theta) = P_{\mathrm{tar}} C_{\mathrm{FR}} \mathrm{FRR}(\theta) + (1 - \mathrm{P_{tar}})\,\mathrm{C_{FA}} \mathrm{FAR}(\theta) \quad (1)$$

P_{tar} is the prior probability of a target speaker, and was set at 0.5. C_{FR} and C_{FA} are the cost of false rejection and false acceptance errors, and were both set equal to 1. θ is the decision threshold. As shown in [30], with these parameters, C_{det} becomes the average of the FAR and the FRR, which is equivalent to the half total error rate (HTER), a frequently used performance metric in biometric identification, e.g., [31]. Prior to evaluating C_{det}, scores were converted into 'calibrated likelihood ratios' [32]: linear calibration parameters w_0 and w_1 were optimized on scores of trials in the range 0–10. The scores S of all trials were then linearly mapped to calibrated likelihood ratios S_{cal}:

$$S_{\mathrm{cal}} = w_0 + w_1 S \quad (2)$$

For each age difference range in Table 2 (greater than 0–1), C_{det} was calculated by thresholding all S_{cal} at LLR = 0, i.e. where $-\log\left(P_{\mathrm{tar}}\left(1 - P_{\mathrm{tar}}\right) C_{\mathrm{FR}} / C_{\mathrm{FA}}\right) = 0$ [33]. The *minimum* detection cost, $C_{\mathrm{det}}^{\mathrm{min}}$, was also obtained, by adjusting the decision threshold θ to minimize C_{det}. These error metrics are shown in Figure 3, along with the EER values from Table 2 for comparison.

In the male case, top of Figure 3, C_{det} increases progressively for the i-vector and GMM-UBM systems. It diverges from the EER after 20 years, indicating that the detection performance is decreasing. $C_{\mathrm{det}}^{\mathrm{min}}$ is close to the EER, and slightly lower above 40 years. However, it too increases progressively with age difference, indicating that even with a well-chosen decision threshold, performance decreases. Calibration therefore becomes more difficult with age-difference. The overall trends are similar in the female case, with variability at the 31–40 age range.

6. Discussion

In this paper, the effect of ageing on the performance of a state-of-the-art i-vector speaker verification system was presented and compared to a classic GMM-UBM approach. An experimental protocol was designed that maximised the number of trials given the TCDSA database.

For male speakers, the i-vector system significantly outperforms the GMM-UBM approach at 'short' age difference ranges (0–1 and 0–10) and overall (over the complete 0–60 range). This result is expected given the additional levels of inter-session compensation present in the i-vector framework. For female speakers however, there is a less significant difference between the i-vector and GMM-UBM approaches (in terms of weighted EER) at 'short' age difference ranges (0–1 and 0–10) and overall (over the 0–60 range). It is unclear why there is not a more marked improvement with the i-vector system, though the 'imbalances' in the TCDSA data are likely a contributing factor.

Unlike the GMM-UBM system, the i-vector system was not optimised for this experiment (by including data similar to the TCDSA database content in development, for example). Despite this, it performs effectively: the EER of the i-vector system for males at the 0–1 range (4.61%) is comparable to the performance of the system in the NIST SRE 2012 common conditions 'CC1' (EER = 5.75%) and 'CC3' (EER = 4.83%) [17]. For females, the EER at the 0–1 range (6.90%) is slightly higher than the common conditions 'CC1' (EER = 4.86%) and 'CC3' (EER

absolute age difference:		0–1	0–10	11–20	21–30	31–40	41–50	51–60
male								
EER %	i-vector	4.61	8.13	10.50	13.83	18.37	22.49	32.74
	GMM-UBM	6.90	12.99	14.14	19.93	26.52	29.83	25.81
	num. tar	2247	9822	4118	3202	1689	819	248
	num. non-tar	4826	25025	24417	20175	13343	10886	5184
female								
EER %	i-vector	9.64	11.62	16.74	19.54	31.99	30.83	24.07
	GMM-UBM	9.62	12.56	19.56	19.44	31.54	23.79	18.92
	num. tar	441	2965	1751	1236	1240	1207	437
	num. non-tar	1734	9104	8592	6937	4486	2280	1059

Table 2: Speaker weighted EERs of i-vector and GMM-UBM systems for trials at increasing absolute age difference ranges. ‘num. tar’ and ‘num. non-tar’ are the number of target and non-target trials respectively.

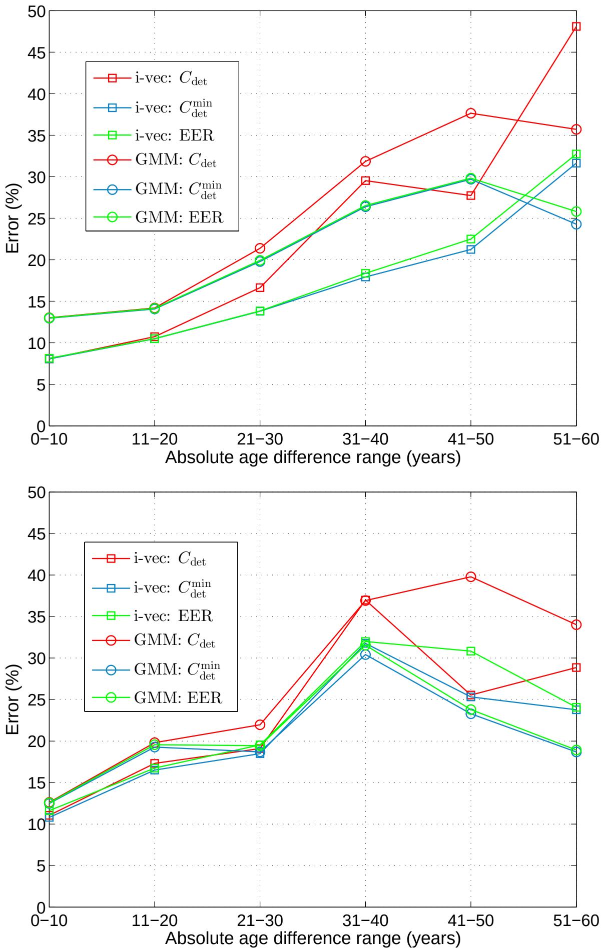

Figure 3: Error metrics for i-vector (‘i-vec’) and GMM-UBM (‘GMM’) systems at increasing absolute age difference ranges, **Top:** Male, **Bottom:** Female

= 4.09%) [17]. This indicates that the i-vector system may not be as well suited to the female content of the TCDSA database, resulting in the greater discrepancy between male and female EER at the 0–60 age range (11.52% vs 16.69%).

The relative change in i-vector performance (in terms of both EER and $C_{\rm det}$) in the presence of ageing is equivalent to that of the GMM-UBM system, demonstrating that long-term ageing variability is not removed by the inter-session compensation applied in the i-vector framework. This therefore motivates dedicated compensation approaches for ageing variability, such as the model-based eigenageing compensation [9] or score-based compensation [8] proposed previously.

Eigenageing compensation was not applied in this paper due to the requirement for an independent set of ageing speakers. With additional data, ageing compensation could also be applied in the PLDA model of the i-vector system. Compensation at a score-level could be achieved by a calibration procedure, treating ageing information as a ‘Quality Measure Function (QMF)’ [32, 34]. The imbalance of the trials across speakers and age-differences make this approach challenging. Ultimately, additional ageing data will be necessary to evaluate these compensation strategies given the experimental protocol adopted in this paper.

As age difference increases, female $C_{\rm det}$ is generally higher in absolute terms, and more rapidly increasing, than male $C_{\rm det}$. A gender-dependent modelling approach is usually taken in speaker verification systems. However, performance is often worse with female speakers, e.g., observed in an evaluation of both NIST SRE 2008 and 2010 tasks [35]. Ageing is a gender-dependent process [1] and thus it may inflate previously existing gender-dependent performance differences. In addition to an ageing compensation approach that is gender-dependent, a front-end that is gender-aware may be necessary to overcome this male/female performance discrepancy.

The large changes to the EER after balancing the contribution from each speaker and each age difference range shows the importance of condition weighting in results analysis. The large spread in the number of per-speaker recordings, and the small number of speakers make the effect particularly significant here. This is likely to be an unavoidable feature of long-term ageing evaluations, due to the nature of the data that is required.

The results of the evaluation in this paper provide benchmark levels of performance on the TCDSA database. Since the database has been made freely available for academic research, this will hopefully stimulate research in the area.

7. Acknowledgements

This research was conducted during a Short Term Scientific Mission (STSM) funded by COST action IC1106

8. References

[1] S. E. Linville, *Vocal Aging*. Canada: Singular, 2001.

[2] P. B. Mueller, "The Aging Voice," *Seminars in Speech and Language*, vol. 18, no. 2, pp. 159–169, 1997.

[3] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in Acoustic Characteristics of the Voice across the Life Span: Measures from Individuals 4-93 Years of Age," *Journal of Speech, Language, and Hearing Research*, vol. 54, pp. 1011–1021, 2011.

[4] P. Torre III and J. A. Barlow, "Age-related changes in acoustic characteristics of adult speech," *Journal of Communication Disorders*, vol. 42, no. 5, pp. 324–333, 2009.

[5] S. Schötz, "Perception, Analysis and Synthesis of Speaker Age," Ph.D. dissertation, 2006, Lund University, Sweden.

[6] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Quality measures in biometric systems," *IEEE Security & Privacy*, vol. 10, no. 6, pp. 52–62, 2012.

[7] A. Lanitis, "A Survey of the Effects of Aging on Biometric Identity Verification," *International Journal of Biometrics*, vol. 2, no. 1, pp. 34–52, 2010.

[8] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language*, vol. 27, no. 5, pp. 1068–1084, 2013.

[9] F. Kelly, N. Brümmer, and N. Harte, "Eigenageing Compensation for Speaker Verification," in *InterSpeech 2013*, Lyon, France, 2013.

[10] F. Kelly and N. Harte, "Auditory detectability of vocal ageing and its effect on forensic automatic speaker recognition," in *InterSpeech 2013*, Lyon, France, 2013.

[11] R. Rhodes, "Changes in the voice across the adult lifespan: formant frequency-based likelihood ratios and ASR performance," in *The International Association of Forensic Phonetics and Acoustics (IAFPA) 2013*, 2013.

[12] ——, "Assessing the strength of non-contemporaneous forensic speech evidence," Ph.D. dissertation, 2012, The University of York.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[14] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation: Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 23, pp. 225–254, 2000.

[15] C. Greenberg, V. Stanford, A. Martin, M. Yadagiri, G. Doddington, J. Godfrey, and J. Hernandez-Cordero, "The 2012 nist speaker recognition evaluation," in *InterSpeech 2013*, 2013.

[16] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[17] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P.-M. Bousquet, E. Khoury, P. L. S. Martinez, J. M. K. Kua, C. You, H. Sun, A. Larcher, P. Rajan, V. Hautamki, C. Hanilci, B. Braithwaite, G.-H. Rosa, S. O. Sadjadi, G. Liu, H. Boril, N. Shokouhi, D. Matrouf, L. El Shafey, P. Mowlaee, J. Epps, T. Thiruvaran, D. Van Leeuwen, B. Ma, H. Li, J.-F. Bonastre, S. Marcel, J. Mason, and E. Ambikairajah, "I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification," in *InterSpeech*, 2013.

[18] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[19] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[20] R. Saeidi and D. van Leeuwen, "The radboud university nijmegen submission to nist sre 2012," in *NIST Speaker Recognition Evaluation Workshop*, 2012.

[21] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[22] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[23] M. McLaren and D. A. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *NIST 2011 SRE Workshop*, 2011.

[24] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Odyssey 2001*, 2001.

[25] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. of ICSLP*, 2006.

[26] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *InterSpeech 2011*, 2011.

[27] F. Kelly, A. Drygajlo, and N. Harte, "Speaker Verification with Long-Term Ageing Data," in *International Conference on Biometrics (ICB) 2012*, New Delhi, India, 2012.

[28] D. A. van Leeuwen, "Overall performance metrics for multi-condition speaker recognition evaluations," in *Proc. Interspeech*. Brighton: ISCA, September 2009, pp. 908–911.

[29] G. Doddington, "The Effect of Target/Non-Target Age Difference on Speaker Recognition Performance," in *Odyssey 2012*, 2012.

[30] M. I. Mandasari, M. Gnther, R. Wallace, R. Saeidi, S. Marcel, and D. A. v. Leeuwen, "Score calibration in face recognition," *IET Biometrics*, 2014, available online 26th February 2014.

[31] K. Kryszczuk and A. Drygajlo, "Improving biometric verification with class-independent quality information," *Signal Processing, IET*, vol. 3, no. 4, pp. 310–321, 2009.

[32] M. Mandasari, R. Saeidi, M. McLaren, and D. van Leeuwen, "Quality measure functions for calibration of speaker recognition system in various duration conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.

[33] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Springer, 2007, vol. 4343.

[34] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. v. Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *ICASSP 2013*, 2013.

[35] S. Cumani, O. Glembek, N. Brummer, E. de Villiers, and P. Laface, "Gender independent discriminative speaker recognition in i-vector space," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4361–4364.