Analysis of mutual duration and noise effects in speaker recognition: benefits of condition-matched cohort selection in score normalization

Andreas Nautsch[⋆], *Rahim Saeidi*[†], *Christian Rathgeb*[⋆], *and Christoph Busch*[⋆]

[⋆]da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
{andreas.nautsch,christian.rathgeb,christoph.busch}@{cased|h-da}.de
[†]Department of Signal Processing and Acoustics, Aalto University, Finland
{rahim.saeidi}@aalto.fi

Abstract

The biometric and forensic performance of automatic speaker recognition systems degrades under noisy and short probe utterance conditions. Score normalization is an effective tool taking into account the mismatch of reference and probe utterances. In an adaptive symmetric score normalization scheme for state-of-the-art i-vector recognition systems, a set of cohort speakers are employed to calculate the mean and variance of impostor scores when compared to reference and probe i-vectors. In dealing with real-life conditions where the quality of audio recordings in test phase does not match enrolment utterance(s) of speakers, we demonstrate the effectiveness of utilizing a condition-matched cohort set for score normalization. The cohort set audio material is shortened and degraded by noise in different reasonable and controlled signal-to-noise ratios according to expected test conditions, yielding in multiple set of cohorts. Further, we propose automatic cohort pre-selection based on modeling each degradation category. For each i-vector, a quality vector is assigned as the posterior probability of degradation classes. The cohort set is then formed by i-vectors representing small KL-divergence of respective quality vectors when compared to reference and probe. Further gains are observed by including this quality vector also into the score calibration.

Index Terms: speaker recognition, AS-norm, duration, SNR, cohort selection, quality characterization

1. Introduction

Biometric speaker recognition becomes more and more valuable for commercial and forensic applications. Thereby, automated recognition systems need to be inter alia duration and noise robust in order to cover a vast broadness of environmental constraints. Literature usually emphasizes on one type of signal degradation so far, and many commercial systems are developed for restricted environments. In this work, mutual duration and noise effects are examined, i.e. sample completeness and ambient noise.

By placing focus on the score post-processing, we examine adaptive symmetric score-normalization (AS-norm) techniques. In practice, a conventional AS-norm is applied [1, 2]. Condition-informed (unconstrained) AS-norm generally increases recognition accuracy by relying on higher-evident comparison statistics, which was shown to be effective on duration conditions in [3]. However, by extending this approach to mutual duration and noise conditions, practical issues will arise in terms of reliable signal-noise-ratio (SNR) estimations: properties of the underlying clean samples will remain unknown to the system.

Thus, we suggest to use unified audio characteristics (UACs), proposed in [4], for cohort pre-selection: sample quality vectors (q-vectors) are derived from condition posterior probabilities. Probe-alike cohort templates are determined by the minimum (symmetric) Kullback-Leibler divergence. Hence, condition-matched cohort sets can be approximated. Thereby, the theoretical framework on using quality measures [5] is extended by the score normalization stage.

Experiments are conducted for five duration and five SNR conditions. SNR conditions stem from two noise sources, in particular: air conditioner (AC) and crowd (CROWD) noise. By degenerating voice samples from the I4U file list of NIST SRE'12 [6], mutual quality and completeness degradation effects are examined on 55 conditions on a state-of-the-art system comprising i-vector features [7, 8] and probabilistic linear discriminant analysis (PLDA) comparison [9, 10, 11]. While in many scenarios reference samples can be captured under very good conditions, probe samples are affected by signal degradation, hence emphasis is put on condition-variable probe samples.

This paper is organized as follows: in Section 2 related work on adaptive symmetric score normalization is briefly summarized and the cohort selection scheme of unconstrained AS-norm is explained. Section 3 links the idea of audio characterization to automatic cohort pre-selection. Experimental evaluations are carried out in Section 4, and conclusions are given in Section 5.

2. Adaptive symmetric score normalization

Score-normalization augments log-likelihood ratio (LLR) scores with additional knowledge of prior observations. Symmetric score normalization (s-norm) considers zero normalization (z-norm) and test normalization (t-norm). In z-norm, scores are normalized by the score distribution of cohort probes with respect to each reference [12]. In t-norm, acoustic effects of probes are addressed, such that score distributions over cohort references are normalized with respect to each probe [13]. State-of-the-Art JFA and i-vector systems utilize either cosine or PLDA comparison. Where cosine comparison based systems, usually consider in-series normalization e.g., zt-norm [13], PLDA based approaches refer to symmetric fusion of z-norm and t-norm, which is known as s-norm [7]. Adaptive variations (AZ, AT, AS), which utilize most competitive sub-cohorts, are considered being more robust by omitting very low cohort scores of these norms [1, 2, 13].

Table 1: *Label scheme for mutual duration and noise conditions.*

Condition	1	2	3	4	5	6	7	8	9	10	11 … 15	16 … 30	31 … 55
Duration	5 s	10 s	20 s	40 s	full	5 s					10 s	20 s … full	5 s … full
Noise	clean					AC							CROWD
SNR						0 dB	5 dB	10 dB	15 dB	20 dB	0 … 20 dB	0 dB … 20 dB	0 dB … 20 dB

2.1. Conventional AS-norm

Conventioned AS-norm (cAS) considers well-enrolled cohort subjects, i.e. clean/full cohort samples are used for extracting cohort templates [1, 2]. Comparison scores are computed, and reference (ref) and probe (prb) i-vectors are scored against all cohort templates using the same comparator as on reference – probe comparison. From each of the two resulting score sets $\mathfrak{R}, \mathfrak{P}$ representing reference – cohort and cohort – probe scores, first and second moment statistics $\mu_{ref}, \sigma_{ref}, \mu_{prb}, \sigma_{prb}$ are derived based on the top-n $\mathfrak{R}, \mathfrak{P}$ scores, respectively. Thereby, the most competitive cohort scores are symmetrically selected with adaptation towards reference and probe features. AS-normalized scores S_{AS} are computed by an averaged symmetric zero-normalization of comparison score S:

$$S_{AS} = \frac{1}{2}\left(\frac{S-\mu_{ref}}{\sigma_{ref}} + \frac{S-\mu_{prb}}{\sigma_{prb}}\right). \tag{1}$$

2.2. Unconstrained AS-norm

Contrary to cAS, unconstrained, condition-informed AS-norm (uAS) takes (easy) quantifiable sample conditions into account in cohort selection, such as duration and noise. The main idea is to achieve reference- or probe-matching conditions for the cohort set [3]. In this terms, uAS seeks probe-alike cohort samples for $\mathfrak{R}$-comparisons and full/clean cohort samples for $\mathfrak{P}$-comparisons (as in cAS) [3]. In terms of Eq. (1), this receipt interprets $\frac{S-\mu_{ref}}{\sigma_{ref}}$ as normalization against successful cohort impostor attempts on references, and $\frac{S-\mu_{prb}}{\sigma_{prb}}$ as normalization against successful probe impostor attempts on cohorts.

Condition-matching cohort selection schemes are expected to not only normalize false matches on references and false non-matches on probes, but also to encounter condition-depending signal degradation.

3. Cohort pre-selection by audio quality

The optimal score normalization needs to utilize cohorts whose audio characteristics correspond best with probe or reference. In order to establish an automated mechanism for extracting reliable audio quality metrics and select an appropriate cohort set, we propose a probabilistic cohort pre-selection scheme based on the unified audio characteristic approach of [4] aiming at posterior probabilities of conditions. The selection scheme favors cohort templates having the lowest relative information divergence to the characteristics of a probe.

3.1. Audio characterization

For the purpose of measuring condition posteriors, single multivariate Gaussian models $\Lambda_i \sim \mathcal{N}(\mu_i, \Sigma), i = 1, \dots, 55$ are trained in original i-vector space. The models have condition-dependent mean vectors μ_i and share a full covariance matrix Σ. Class-dependent means are estimated using i-vectors from respective quality condition and Σ is estimated by pooling all the i-vectors. The resulting vector of posterior probabilities for an i-vector $\boldsymbol{w}$ represents a condition quality vector (q-vector) $\boldsymbol{q}$, where:

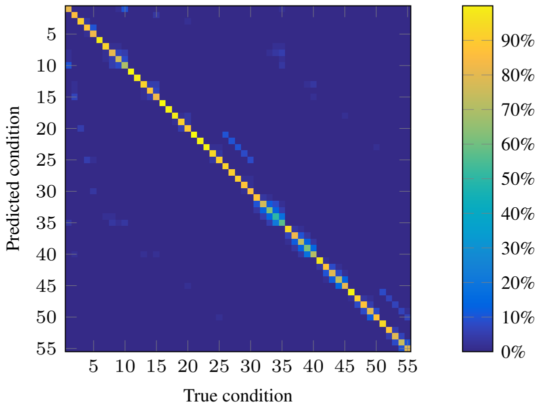


Figure 1: *Condition confusion matrix on q-vector max-posterior classification* $\max P(\Lambda_{1,\dots,55}|\boldsymbol{w})$.

$$q(i) = \frac{P(\boldsymbol{w}\,|\,\Lambda_i)}{\sum_{i=1}^{55} P(\boldsymbol{w}\,|\,\Lambda_i)}. \tag{2}$$

All templates (ref, prb, cohort) are extended to a pair of an i-vector and a corresponding q-vector.

Figure 1 depicts the confusion matrix among all conditions, where the condition indexes are defined in Table 1. While conditions 31 – 40, comprising the highest signal degradation of 5 s/CROWD and 10 s/CROWD, are more likely to be confused with other 31 – 40 conditions (up tp 51% misclassification rate), the vast majority of conditions are far more well-classified, i.e. with less than 20% mis-classification and up to 99.6% correct classification rates. On AC and CROWD noises, 10% of 40 s/noisy conditions are recognized as their full/noisy condition equivalents having similar SNR levels.

3.2. Cohort selection criterion

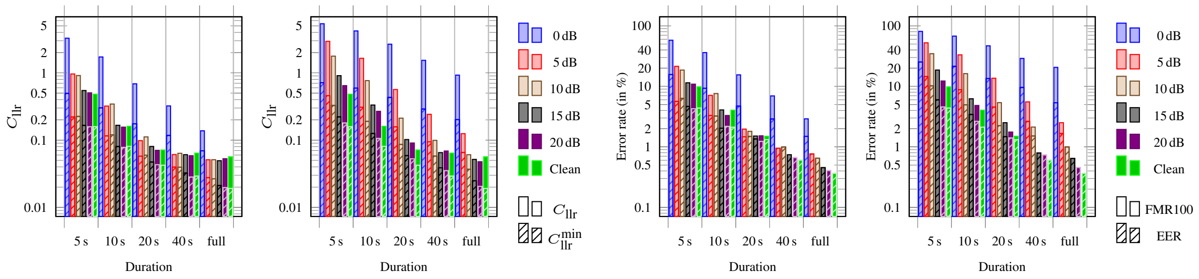
While conditions can be classified by the maximum posterior probability, the cohort selection requires a similarity metric to find *close* audio segments in sense of q-vector. Inspired by [14], we propose the symmetric Kullback-Leibler divergence D_{symKL} of two q-vectors $\boldsymbol{q_a}, \boldsymbol{q_b}$ for pre-selecting cohorts:

$$D_{\text{symKL}}(\boldsymbol{q_a}||\boldsymbol{q_b}) = \frac{1}{2}\sum_{i=1}^{55} \boldsymbol{q_a}(i)\log\frac{\boldsymbol{q_a}(i)}{\boldsymbol{q_b}(i)} + \boldsymbol{q_b}(i)\log\frac{\boldsymbol{q_b}(i)}{\boldsymbol{q_a}(i)}. \tag{3}$$

The closest top-c cohort q-vectors are denoted by $\min D_{\text{symKL}}$.

4. Experimental results

Condition-dependent sample versions were created from long-duration and clean samples of the I4U file list prepared for sites participating in NIST SRE'12 [6] by truncation into duration



(a) AC noise (b) CROWD noise (c) AC noise (d) CROWD noise

Figure 2: *Baseline performance on I4U eval-set affected by mutual signal degradation, no calibration.*

groups of 5 s, 10 s, 20 s, 40 s, and full (original duration) as in [15], and by applying AC and CROWD noise using FaNT, such that noise groups of 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and clean (original SNR) were established. In total 55 conditions were examined (see Table 1).

4.1. Experimental setup

Stabilized weighted linear prediction (SWLP) [16] is employed for robust spectrum estimation after enhancing the audio using maximum-likelihood short-time spectral amplitude (MLSTSA) [17]. The rest of the front-end processing is similar to our earlier work in [18, 19]. Raw i-vectors were drawn from samples after Voice Activity Detection (VAD). The VAD labels from clean condition are then applied to corresponding noise versions. We assume perfect VAD for this experiments in order to exclude undesirable effects rising from VAD shortcomings in low SNRs. Full/clean probe samples of the I4U development set (dev-set) and evaluation set (eval-set) of all male subjects were modified condition-dependently.

Then, all reference samples of the dev-set were modified condition-dependently in order to serve fair i-vector processing: discriminant spherical space projection was performed by an LDA dimension reduction [20] from 400 to 200 dimensions, within-class-covariance-normalization (WCCN) [19] and length-normalization [10]. For the sake of tractability of analysis, we experiment with only male speakers data. The i-vectors are compared by PLDA [10] with 200 speaker factors. PLDA is trained in a multi-condition pooled fashion as in [21].

4.2. Evaluation criteria

The biometric performance is reported in accordance to the ISO/IEC IS 19795-1 [22] by the Equal-Error-Rate (EER), and the False Non-Match Rate (FNMR) at a 1% False Match Rate (FMR100). As an application-independent performance metric, we emphasize on the minimum cost of log-likelihood ratio (LLR) scores $C_{\text{llr}}^{\text{min}}$, which represents the generalized empirical cross-entropy of genuine and impostor LLRs with respect to Bayesian thresholds $\eta \in (-\infty, \infty)$ assuming well-calibrated systems [23, 24]. The actual C_{llr} [24] is computed over genuine and impostor scores S_G, S_I by:

$$C_{\text{llr}} = \frac{\sum_{g \in S_G} \text{ld}(1 + \frac{1}{e^g})}{2\,|S_G|} + \frac{\sum_{i \in S_I} \text{ld}(1 + e^i)}{2\,|S_I|}. \tag{4}$$

4.3. Baseline results

Figure 2 shows the performance of a state-of-the-art baseline system without score normalization nor calibration. In general, CROWD noise causes higher performance deterioration than AC noise. Longer duration and lower SNRs lead to higher performance, and in the vast majority of conditions, clean/full outperforms other conditions with 0.019 $C_{\text{llr}}^{\text{min}}$ and 0.4% EER (as expected). By focusing on the effect of noise level on $C_{\text{llr}}^{\text{min}}$, 0 dB on AC and 0 dB and 5 dB on CROWD are causing high signal degradation leading to significantly worse performance, while the performance of 20 dB and clean is very similar on AC across duration conditions. By shifting the focus on duration effects, EER and $C_{\text{llr}}^{\text{min}}$ performance linearly depend on the log-duration as observed by [18], and mutual effects appear as a linear combination of log-duration and log-SNR impacts.

4.4. Analysis: i-vector pool mean shift

In order to measure i-vector property changes by signal degradation, we examine the i-vector condition means, raising the question whether cross-condition i-vectors share the same mean or not. Contrary to [3], where i-vectors were element-wise tested for shared mean dimensions by Student t-test, in this work we consider vector space mean among the i-vector condition pools by utilizing the generalized, multi-variate Student t-test, in particular: the Hotelling's T-squared statistic [25, 26]. In the according statistic test for population-independent means, H_0 states i-vectors sharing the same mean among conditions, H_1 states different i-vector means by assuming equal covariances. The test value of the generalized Student's t-test t^2 utilizes the averaged scatter of both populations $\mathbf{W}$ and is defined as:

$$t^2 = \frac{n_x\, n_y}{n_x + n_y}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})'\mathbf{W}^{-1}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}), \textit{with:}$$

$$\mathbf{W} = \frac{\sum_{i=1}^{n_x} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' + \sum_{i=1}^{n_y} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'}{n_x + n_y - 2},$$

$$\bar{\boldsymbol{x}} = \frac{\sum_{i=1}^{n_x} \boldsymbol{x}_i}{n_x}, \qquad \bar{\boldsymbol{y}} = \frac{\sum_{i=1}^{n_y} \boldsymbol{y}_i}{n_y}, \tag{5}$$

where n_x, n_y are the number of observations on D-multivariate data sets $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. In this experimental setup, D equals 200. P-values are estimated by the cumulative distribution function F of χ^2 distributions [25, 26]:

$$t^2 \sim \chi^2_D,$$

$$p = 1 - F_{\chi^2_D}(t^2). \tag{6}$$

Figure 3 illustrates observed test values between all 55 conditions; p-values will result either as exactly one on χ^2 scores of zero, or as p-values lower than $10^{-13} \approx 0$ indicating high significance. Only same-condition tests result in zero χ^2 scores. Hence, all cross-condition mean shifts are highly significant.

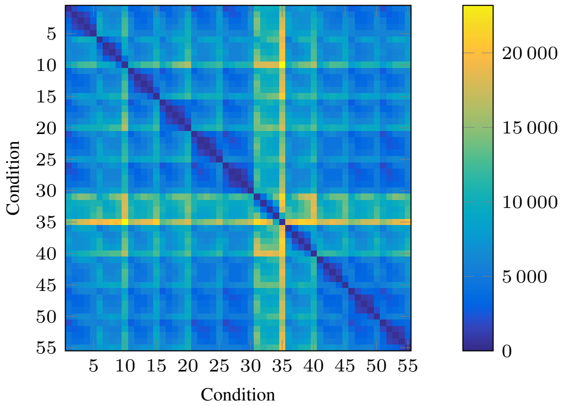


Figure 3: *Multi-variate Student's t-test: comparison of χ^2 scores examining i-vector mean-similarity among conditions.*

4.5. Oracle versus automatic cohorts pre-selection

By taking information of other comparisons into account, AS-norm can compensate subject and condition-dependent variances on score domain. Examining $C_{\text{llr}}^{\text{min}}$ performance, uAS outperformed the baseline (and cAS) among the vast majority of conditions with relative gains up to 8.2% in $C_{\text{llr}}^{\text{min}}$, 15.9% in EER, and 23.4% in FMR100. The cohort size in terms of top-n selection size, however, had no impact on this metrics, making a cohort size of 50 interesting for least-effort concerns. We also examined a cohort selection scheme seeking reference-alike $\mathfrak{R}$ cohorts and probe-alike $\mathfrak{P}$ cohorts, however no sufficient gains to cAS were observed, confirming the uAS approach.

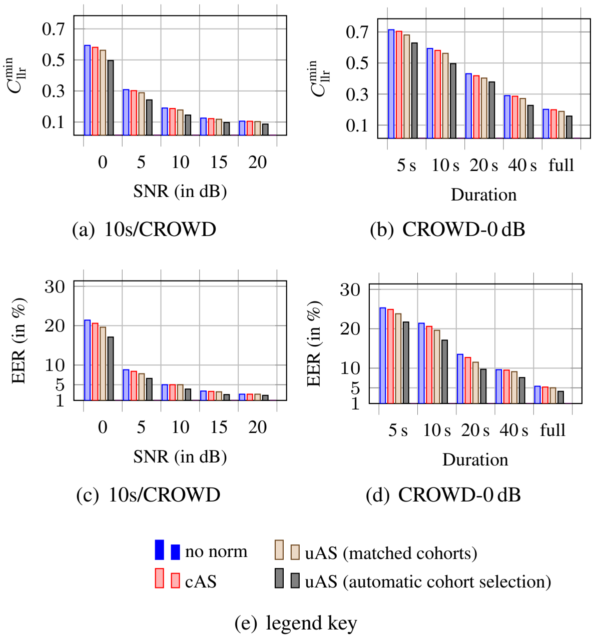
Aiming at mutual high-degradation conditions, Figure 4 compares AS-norms by SNR levels on 10s/CROWD and by duration groups on CROWD-0 dB: the proposed cohort selection significantly outperforms all other systems in $C_{\text{llr}}^{\text{min}}$ and EER including oracle cohort selection of uAS proving the soundness of quality based cohort selection.

Figure 5 illustrates which conditions and cohort subjects were considered in pre-selection: cohorts having similar noise source, duration and SNR level are favored, while the vast majority of other conditions is not considered even in a single cohort speaker. The most amount of cohort templates are selected from conditions 36 – 38 (10s/CROWD-0 – 10 dB), then from conditions 39 and 40 completing the block of SNR levels in 10s/CROWD conditions. Noise source impacts reveal from condition 2 and 11 selections representing 10s/clean and 10s/AC-0 dB, respectively. Duration impacts reveal from selections of conditions 31 –35 denoting 5s/CROWD noise conditions. This pattern is also observed on increasing duration, where much more cohort speakers are considered among duration and noise similar conditions by longer probe durations.

Inspired by [5], by including q-vector information into calibration stage in terms of:

$$S' = w_0 + w_1\, S + w_2 \cos(\boldsymbol{q}_{\mathit{ref}}, \boldsymbol{q}_{\mathit{prb}})\, S, \qquad (7)$$

with weights w_0, w_1, w_2, S' stating the calibrated score, and S denoting the score of the proposed uAS method, a marginal gain is observed in system performance. Eq. (7) assumes score impact factors based on the quality gap between reference and probe i-vectors which is represented by the angle between reference and probe q-vectors $\boldsymbol{q}_{\mathit{ref}}$, $\boldsymbol{q}_{\mathit{prb}}$. We have left finding more efficient techniques for including q-vectors in calibration stage for further research.



(a) 10s/CROWD

(b) CROWD-0 dB

(c) 10s/CROWD

(d) CROWD-0 dB

(e) legend key

Figure 4: *C_{llr}^{min} and EER comparison of conventional AS-norm to oracle cohorts and the proposed pre-selection by q-vectors in extreme conditions.*

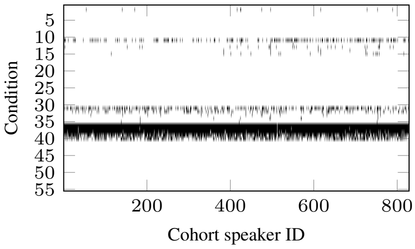


Figure 5: *Pre-selected cohort subjects and conditions on 10s/CROWD-0 dB (condition 36) by unique selection (black).*

5. Conclusion

Mutual duration and noise effects severely effect speaker recognition concerning sample completeness and quality. Condition-informed (unconstrained) AS-norm robustly improves biometric and forensic performances, but it is clearly not capable of reaching the performance on full/clean samples. However, by quality-based cohort pre-selection instead of relying on oracle cohort sets, significant gains in biometric and forensic performance are yielded, such that this approach seems also promising for other, similar issues, such as domain shift compensation.

6. Acknowledgment

We would like to thank the I4U consortium for database sharing. This work has been partially funded by the Center for Advanced Security Research Darmstadt (CASED), the Hesse government (project no. 467/15-09, BioMobile) and the Academy of Finland (project no. 256961 and 284671).

7. References

[1] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of Speaker Recognition Approaches for Real Applications," in *Interspeech*. ISCA, 2011.

[2] D. Colibro, C. Vair, K. Farrell, N. Krause, G. Karvitsky, S. Cumani, and P. Laface, "Nuance - Politecnico di Torino (NPT) System Description for NIST 2012 Speaker Recognition Evaluation," in *Proc. NIST SRE'12 workshop*, 2012.

[3] A. Nautsch, C. Rathgeb, C. Busch, H. Reininger, and K. Kasper, "Towards Duration Invariance of i-Vector-based Adaptive Score Normalization," in *Odyssey 2014: The Speaker and Language Recognition Workshop*. ISCA, 2014.

[4] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, "A Unified Approach for Audio Characterization and its Application to Speaker Recognition," in *Odyssey 2012 – The Speaker and Language Recognition Workshop*. ISCA, 2012.

[5] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "On the use of quality measures for text-independent speaker recognition," in *Odyssey 2004 – The Speaker and Language Recognition Workshop*. ISCA, 2004.

[6] R. Saeidi, K. A. Lee, T. Kinnunen et al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Interspeech*. ISCA, 2013.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech and Language Processing, IEEE Transactions on*, 2010.

[8] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge," in *Odyssey 2014: The Speaker and Language Recognition Workshop*. ISCA, 2014.

[9] S. J. Prince, *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012. [Online]. Available: http://www.computervisionmodels.com/

[10] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*. ISCA, 2011.

[11] S. Cumani and P. Laface, "Generative pairwise models for speaker recognition," in *Odyssey 2014: The Speaker and Language Recognition Workshop*. ISCA, 2014.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," in *Conversational Speech, Digital Signal Processing*, 2000.

[13] D. E. Sturim and D. A. Reynolds, "Speaker Adaptive Cohort Selection for Tnorm in Text-independent Speaker Verification," in *International Conference on Acoustic, Speech and Signal Processing*. IEEE, 2005.

[14] G. Aradilla, J. Vepa, and H. Bourlard, "Using Posterior-Based Features in Template Matching for Speech Recognition," in *Interspeech*. ISCA, 2006.

[15] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions," *Transactions on Audio, Speech, and Language Processing, IEEE Transactions on*, 2013.

[16] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, 2010.

[17] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1980.

[18] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration Mismatch Compensation for i-Vector based Speaker Recognition Systems," in *International Conference on Audio, Speech and Signal Processing*. IEEE, 2013.

[19] R. Saeidi and D. A. van Leeuwen, "The Radboud University Nijmegen submission to NIST SRE-2012," in *Proc. NIST SRE'12 workshop*, 2012.

[20] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction: A comparative review," Tilburg University, Tech. Rep., 2009, tiCC-TR 2009-005.

[21] 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 *International Conference on Acoustic, Speech and Signal Processing*. IEEE, 2012.

[22] ISO/IEC, "Information technology – Biometric performance testing and reporting – Part 1: Principles and framework," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 19795-1:2006, 2011.

[23] D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy Analysis of the Information in Forensic Speaker Recognition," in *Odyssey 2008: The Speaker and Language Recognition Workshop*. ISCA, 2008.

[24] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," in *Odyssey 2006: The Speaker and Language Recognition Workshop*. IEEE, 2006.

[25] H. Hotelling, "The generalization of Student's ratio," *Annals of Mathematical Statistics*, 1931.

[26] A. Trujillo-Ortiz and R. Hernandez-Walls, "HotellingT2: Hotelling T-Squared testing procedures for multivariate tests. A MATLAB file." 2002. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/2844-hotellingt2