ENTROPY ANALYSIS OF I-VECTOR FEATURE SPACES IN DURATION-SENSITIVE SPEAKER RECOGNITION

Andreas Nautsch[⋆] *Christian Rathgeb*[⋆] *Rahim Saeidi*[†] *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 vast majority of speaker recognition cross-entropy evaluations are focused on score domain. By examining the generalized relative distance between genuine and impostor subspaces, biometric characteristics become comparable to other authentication approaches. In this paper we demonstrate that the i-vector feature space's biometric information measured by relative entropy is comparable to e.g., knowledge-based mechanisms or face recognition.

Examining NIST SRE 2004-2010 corpora, short samples of e.g, 5 seconds duration, comprise already 127 bits in a text-independent scenario. Further, the vast majority of short samples does not fall below 50% of the biometric information of samples having a duration of more than 40 seconds. The generalized i-vector feature space entropy of long samples corresponds to 182.1 bits, and the highest lower entropy bound of a subject was observed at 471.6 bits.

Index Terms— biometric information, relative entropy, speaker recognition, i-vector, duration

1. INTRODUCTION

Forensic and industrial applications are based on evidence or security level of an automated recognition system [1]. Additionally, information about the underlying feature space is required [2, 3], in which a feature vector represents the biometric characteristic of a subject [4]. The work in [1] was focused on empirical cross entropy analysis across different recognition systems and in this work we analyze utterance duration effect on feature space entropy. By measuring the relative entropy of one subject's subspace compared to the generalized space of all other subjects, the biometric information within a given feature space can be reported [2]. In comparison, the entropy of passwords or PINs $H(\text{string})$ can be computed by the string length L and the number of different string characters that can occur, N [5]: $H(\text{string}) = L \log_2 N$. Thus, 4-digit PINs have an entropy of 13.3 bits. Compared to more secure passwords having at least 128 bits including special characters (printable & extended ASCII codes $N = 224$) users need to remember passwords having $L = 17$ characters. Further, by emphasizing high-evidence or high-secure systems, it is further of interest to know, when collisions will occur in the best case. This can be directly derived from a feature space's entropy $H(\text{space})$ as the probability p_{col} [6]:

$$p_{\text{col}}(\text{space}) = 2^{-H(\text{space})}, \tag{1}$$

where $p_{\text{col}}(\text{PIN}) \approx 1 \times 10^{-4}, p_{\text{col}}(\text{password}) \approx 3 \times 10^{-39}$ for 4-digit PINs and secure passwords, respectively. Compared to passwords, biometric systems are user-friendly and avoid problems of remembering passwords, since the key is the subject itself in biologic or behavioral terms [7].

We will show that the relative entropy in speaker recognition can be considered to be roughly equal to 128 bits-strong passwords on short samples having less than 20 seconds, and to be much more higher on longer samples. This paper will contribute an evaluation of the biometric information of state-of-the-art speaker recognition systems with respect to probe sample duration, such that speaker recognition systems become more competitive to other authentication methods. We place focus on the analysis of the identity vector (i-vector) feature space [8], which has become state-of-the-art in speaker recognition [9].

2. RELATED WORK

Recent speaker recognition approaches rely on i-vector based speaker representations, which represent the characteristic speaker offsets from an Universal Background Model (UBM), which models the distribution of acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) [10]. Thereby, concatenated UBM component's mean vectors form a *supervector* $\vec{\mu}_{\text{UBM}}$, which is the base of a *Front-End* [8] speech feature space, where a speaker supervector $\mathbf{s}$ is decomposed by a total variability matrix $\mathbf{T}$ into a lower-dimensional and higher-discriminant i-vector $\vec{i}$ as an offset to $\vec{\mu}_{\text{UBM}}$:

$$\mathbf{s} = \vec{\mu}_{\text{UBM}} + \mathbf{T}\,\vec{i}. \tag{2}$$

The total variability matrix is trained on a development set using an expectation maximization algorithm [8, 11]. In order to compare samples of varying duration and to achieve i-vectors with un-correlated vector elements, a spherical projection is performed on the i-vectors by applying whitening transform and length-normalization [12, 13]. As a matter of fact, length-normalized i-vectors mostly follow Gaussian distribution [12]. State-of-the-art i-vector comparators belong to the Probabilistic Linear Discriminant Analysis (PLDA) family [13, 14], where it has been also widely utilized in 2014 NIST i-vector machine learning challenge [9].

Since speakers are separated using statistical models, the biometric performance and evidence strength of recognition systems strongly depends on the reference and probe i-vector's significance [15, 16]. The sample duration has a vast influence on the significance of i-vector features, because this features depend on UBM statistics accumulated over-time, namely zero and first order Baum-Welch statistics, which are sparse on short-term signals [11].

Recent work on compensating duration effects emphasized e.g., in the score, comparator, or i-vector extractor domain: in [16, 17] duration-sensitive score calibration and normalization schemes are applied. In [15, 18] are PLDA comparators trained with respect to certain duration groups. Elaborated i-vector extractors [19] refer to Vector Taylor Series (VTS) expansions of the Baum-Welch statistics. While most of the compensation approaches rely on the same feature space, VTS-related research seek new feature spaces, which promise consistent biometric information independent of a sample's duration. In this paper, we place focus on common MFCC-based i-vectors, where we expect gains by increasing duration, because the correlation of i-vector length, speech duration and performance is an expected phenomenon based on the prior distribution of i-vectors as shown in [12, 15].

3. MEASURING FEATURE SPACE ENTROPY

Estimations for biometric information were done inter alia by Ratha et al. [6], Daugman [3], and Adler et al. [2]. Adler et al. also referred to the biometric information as a measurement for the biometric *uniqueness*. The approaches rely on collision estimations by brute force, estimating the number of independent bits on binarized feature vectors, and the relative entropy between genuine and impostor sub-spaces.

Ratha et al. [6] looked for the probability of guessing the features in random. For fingerprints, they evaluated the total number of possible variations for K minutiae locations, m minutiae, and d number of minutiae orientations, such that they formulated the collision probability as $1/\left(\binom{K}{m} d^m\right)$, from which entropy can be measured using Eq. (1). This approach address the robustness of a feature space on brute-force attacks rather than it's ability to distinguish between subjects.

Daugman [3] analyzed binary iris features, on which the Hamming distance is used for comparing all subjects of a database to each other. He related the score distribution to a Bernoulli-Experiment having $N = \mu(1-\mu)/\sigma^2$ degrees-of-freedom, where μ is the observed Hamming distance mean value and σ^2 is the variance, respectively. A feature space's entropy is referred to by N, which represents the amount of coin tosses needed for a feature space collision. This method describes the unique feature space elements of a binary feature space. On this method, Adler et al. [2] argue that the question of *to what extent are biometric characteristics unique* needs to be more addressed, than the uniqueness that is provided on the feature space elements.

Adler et al. [2] introduced a measurement for *biometric information*, that addresses the inter-subjects information of features $\mathbf{x}$, which is measured by the Kullback-Leibler divergence $D(p||q)$ of the intra-subject distribution $p(\mathbf{x})$ and the inter-subject distribution $q(\mathbf{x})$, and represents the needed extra information (in bits) to represent $p(\mathbf{x})$ with respect to $q(\mathbf{x})$:

$$D(p||q) = \int_{\mathbf{x}=-\infty}^{\infty} p(\mathbf{x}) \log_2 \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}. \tag{3}$$

The p distribution represents a subject's feature sub-space, while the q distribution represents the feature space of all other subjects. It is assumed that p and q follow a Gaussian distribution with parameters $p(\mathbf{x}) \sim \mathcal{N}(\vec{\mu}_p, \mathbf{\Sigma}_p)$ and $q(\mathbf{x}) \sim \mathcal{N}(\vec{\mu}_q, \mathbf{\Sigma}_q)$, respectively. By using the Gaussian model, the Kullback-Leibler divergence represents a lower bound to the estimated relative entropy and Eq. (3) can be formulated as [2]:

$$D(p||q) = k(\lambda + \mathit{trace}((\mathbf{\Sigma}_p + \mathbf{T})\mathbf{\Sigma}_q^{-1} - \mathcal{I})), \tag{4}$$

$$\text{with } k = \log_2\sqrt{e}, \lambda = \ln\frac{|\mathbf{\Sigma}_q|}{|\mathbf{\Sigma}_p|}, \mathbf{T} = (\vec{\mu}_p - \vec{\mu}_q)^t(\vec{\mu}_p - \vec{\mu}_q).$$

The relative sub-space entropy $H(p)$ is computed by the average relative subject entropy. Fig. 1 illustrates how the Gaussian model acts as an lower bound compared to a more sophisticated model, i.e. a Gaussian Mixture Model (GMM) over q on Gaussian-distributed exemplary data.

In order to estimate each subject's relative entropy significantly, Adler et al. [2] refer to two regularization approaches:

a) Regularization for degenerated features: usually high-dimensional feature spaces are extracted from samples e.g., with $F = 400$ dimensions, while the analyzed database may only contain a couple of samples per subject e.g., $N_p = 10$. In order to significantly estimate entropy, the feature space is transformed into a G-dimensional space by the Principal Component Analysis (PCA), where $G \leq F$. The PCA is performed on the q distribution covariance by Singular Value Decomposition (SVD), since the q distribution is much more accurately estimated that the p distribution, such that:

$$\mathbf{U}\mathbf{S}_q\mathbf{V}^{\mathbf{t}} = \mathit{svd}(\mathbf{\Sigma}_q). \tag{5}$$

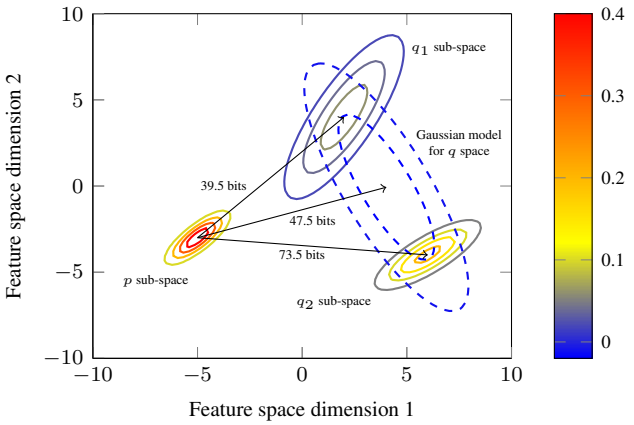

Fig. 1. Estimating the lower bound of relative entropy by using a generalizing Gaussian model compared to a more detailed GMM on exemplary data. The single-Gaussian model estimates a lower bound of 47.5 bits, while the GMM's estimation is calculated by the mean of each GMM component's relative entropy, i.e. as the average distance of the sub-spaces of subjects q_1, q_2 to p with 39.5 bits and 73.5 bits, respectively: *GMM-mean*$(q_1, q_2) = 56.5$ bits.

The matrices $\mathbf{U}, \mathbf{S}_q, \mathbf{V}$ are truncated to the dimension G according to an adaptive impact threshold considering the PCA-impact of the first element $10^{-10}[\mathbf{S}_q]_{1,1}$. Elements are truncated at $[\mathbf{S}_q]_{j,j} < 10^{-10}[\mathbf{S}_q]_{1,1}$. Then, the subject's PCA feature space covariance $\mathbf{S}_p$ is computed by:

$$\mathbf{S}_p = \mathbf{U}^t \mathbf{\Sigma}_p \mathbf{V}, \tag{6}$$

and the Kullback-Leibler divergence is restated as:

$$D(p||q) = k(\nu + \mathit{trace}(\mathbf{U}((\mathbf{S}_p + \mathbf{S}_t)\mathbf{S}_q^{-1}) - \mathcal{I})\mathbf{V}^t), \tag{7}$$

$$\text{with} \quad \nu = \ln \frac{|\mathbf{S}_q|}{|\mathbf{S}_p|} \quad \mathbf{S}_t = \mathbf{U}^t \mathbf{T} \mathbf{V}.$$

b) Regularization for insufficient data: given N_p samples, covariance estimations on $G \geq N_p$ will lead to singular $\mathbf{\Sigma}_p$ and let the entropy diverge to ∞. In order to avoid ill-disposed $\mathbf{\Sigma}_p$, non-diagonal elements $[\mathbf{\Sigma}_p]_{i,j}$ are set to zero at $i, j \geq N_p$ e.g., on $N_p = 10$ all non-diagonal covariances with column or row indexes $i, j \geq 10$ are zeroed, while the diagonal variances remain.

Since this regularization scheme needs to be extended, the truncated matrix derived from a positive finite (covariance) matrix is not necessarily positive finite as well. Adler et al. [2] referred to a database, on which 16 samples are distributed for each subject, such that covariance estimations are much more confidential compared to the case of varying sample amounts per subject with $N_p \leq 10$. Thus, we extended the regularization scheme by:

c) Regularization for ill-conditioned PCA covariances: non-diagonal elements $[\mathbf{\Sigma}_p]_{i,j}$ are iteratively zeroed until $\mathbf{\Sigma}_p$ is positive finite.

d) Regularization for insufficient sample amount: mean and covariance estimations need to be estimated from a proper amount of samples, which can be variable in databases. Facing the properties of most databases, *proper* is denoted, such that only subjects are examined, which have at least $N_p = 10$ samples.

4. DATA ANALYSIS & EXPERIMENTS

For experiments in this paper, we used the NIST SRE 2004-2010 corpora. The evaluation protocol follows I4U file lists [20] and their extended version for studying the speech duration effect in [15]. The system architecture and settings of i-vector extractor used in this study can be found in [21]. The database contains 551 female and 425 male subjects having at least 10 samples, respectively. Subject-disjunct development and evaluation subsets are separated into female and male template and probe data sets, respectively. While both template sets only contain full i-vectors, the probe set contains truncated i-vectors of the duration groups 5, 10, 20, 40 seconds and full (>40 s) stemming from the same sample.

For analytic purposes of estimating the biometric information of state-of-the-art speaker recognition in a duration-sensitive manner, we compared duration-variable p subspaces with full-duration q spaces simulating the automatic recognition case, in which full reference i-vectors are compared to probe i-vectors of all duration groups. Fig. 2 and Tab. 1 compare the relative entropy among the duration scenarios (full-vs-5/10/20/40/full), and show correlations to the biometric and score cross-entropy performance of a corresponding PLDA comparator with 400 speaker factors. 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 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 [1, 23].

In general, biometric information increases by duration, which results in better speaker verification performances and lower $C_{\text{llr}}^{\text{min}}$. This behavior is expected, since i-vectors gain

Table 1. Relative entropy and performance comparison of mixed gender PLDA recognition.

Duration group	Entropy (in bits) μ	σ	min	max	PLDA (400) EER	FMR100	$C_{\text{llr}}^{\text{min}}$
full-5	127.2	24.0	71.5	226.6	17.0%	66.7%	0.529
full-10	124.3	28.1	65.0	254.8	8.7%	31.6%	0.296
full-20	135.5	35.3	63.2	319.0	4.1%	9.8%	0.147
full-40	155.0	43.1	71.1	421.9	2.1%	3.2%	0.078
full-full	182.1	50.0	88.7	471.6	1.7%	2.1%	0.069

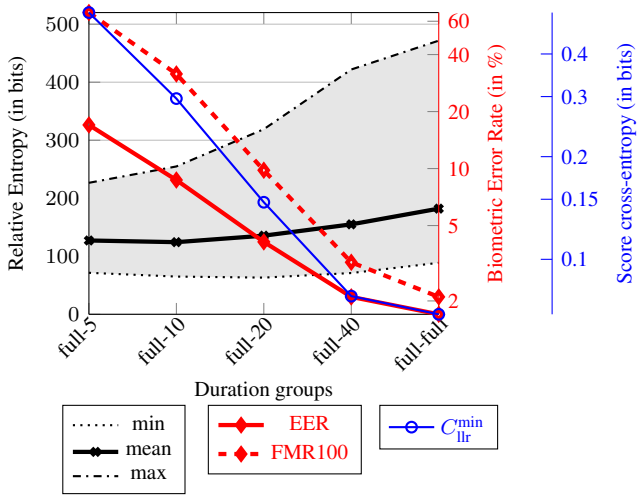

Fig. 2. Comparison of feature and score domain relative entropy along with speaker recognition performance among the duration groups in common recognition scenario (full vs. variable duration) using PLDA with 400 speaker factors on mixed gender.

more significance by duration. The standard deviation of the subject-wise relative entropy increases as well as the maximum entropy. According to the experiments in this paper we observe that the lowest entropy can be estimated to be as low as 63.2 bits fort short duration, and 88.7 bits for full segments. The exact numbers of entropy for different system set-ups could be different, but it is deemed that the trend would be consistent. Where the trend of minimum entropy is higher than the fused face recognition feature space's entropy referred to by Adler et al. [2] with 46.9 bits. The mean of the calculated entropy shows a minimum of 124.3 bits for the full-10 condition, which reveals that even short speech samples can compete with 128 bit-strong passwords in terms of feature space entropy. The biometric information of full segments yields the highest mean entropy value of 182.1 bits. In gender-dependent analysis, we obtained similar results, where the highest relative entropy on the female and male sub-sets for full segments exceeded 300 bits and 400 bits, respectively.

In order to provide more detailed information about the actual respective subject's relative entropy, Fig. 3 visualizes the duration-based accumulation of relative entropy by each subject, in which the relative full-full entropy normalizes the relative entropy of all duration conditions. Besides a few outliers having more biometric information on shorter samples, the vast majority of all relative entropies is within 50 – 100% of the subject-according full-full entropy. Where the subject discrimination in terms of a subject's relative entropy accumulates by increasing duration. However, in comparison to other duration conditions, relative entropy of the full-5 condition is more widely distributed, and partly reach full-full level. Further, full-10/20/40 relative entropy values accumulate continuously, while there is a gain from full-40 to full-full among the vast majority of all subjects.

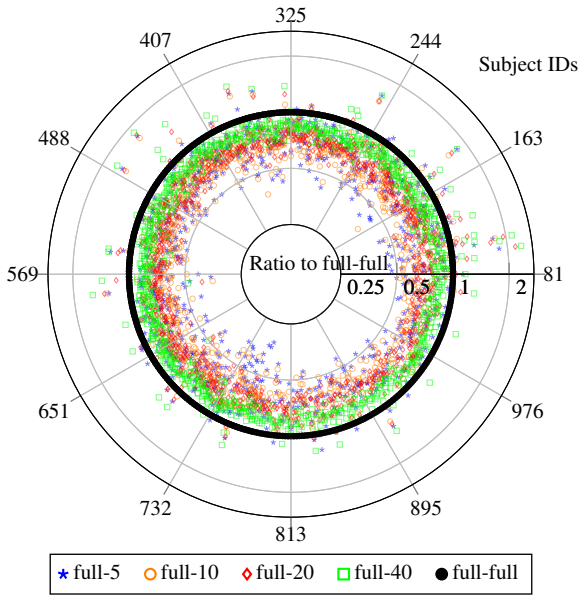

Fig. 3. Speaker sub-space accumulation by duration: relative entropy is subject-wise normalized by the according full-full entropy, such that the accumulation is logarithmic visualized by ratios, i.e. all full segment ratios are 1 perishing actual entropy value comparisons.

5. CONCLUSION

In contrast to existing literature on entropy in speaker recognition, which merely focuses on the score level e.g. [1, 9], this work emphasizes on the feature level. It is demonstrated that current speaker recognition feature spaces reach the relative entropy level of 128 bits-strong passwords already at 20 seconds of speech, where the recognition performance is acceptable. The generalized collision probability of i-vector based speaker recognition can be estimated as $p_{\text{col}}(\text{voice}_{127.2\,\text{bit}}) \approx 5 \times 10^{-39}$ for short samples and $p_{\text{col}}(\text{voice}_{182.1\,\text{bit}}) \approx 2 \times 10^{-55}$ for long samples, respectively, i.e. automated speaker recognition is viable instrument for forensic investigations. From an industrial perspective, voice is found to be a suitable biometric characteristic for user-friendly high-security commercial authentication mechanism, e.g. e-banking. Further gains are expected by fusing i-vectors stemming from different speech signal features.

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) and Academy of Finland (project no. 256961 and 284671).

6. REFERENCES

[1] D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy Analysis of the Information in Forensic Speaker Recognition," in *IEEE Odyssey*, 2008.

[2] A. Adler, R. Youmaran, and S. Loyka, "Towards a Measure of Biometric Information," *Canadian Conference on Electrical and Computer Engineering*, 2006.

[3] J. Daugman, "Probing the Uniqueness and Randomness of IrisCodes: Results From 200 Billion Iris Pair Comparisons," *Proceedings of the IEEE*, 2006.

[4] ISO/IEC, "Information technology — Vocabulary — Part 37: Biometrics," ISO/IEC 2382-37:2012, JTC 1/SC 37, Geneva, Switzerland, 2012, Harmonised biometric vocabulary.

[5] W. E. Burr, D. F. Dodson, and W. T. Polk, "Electronic authentication guideline, recommendations of the national institute of standards and technology, information security," Tech. Rep., National Institute of Standards and Technology (NIST), 2006.

[6] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, 2001.

[7] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

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

[9] 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 *ISCA Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.

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

[11] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep., Centre de recherche informatique de Montréal (CRIM), 2005.

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

[13] P.-M. Bousquet, J.-F. Bonastre, and D. Matrouf, "Identify the Benefits of the Different Steps in an i-Vector Based Speaker Verification System," in *Iberoamerican Congress on Pattern Recognition*, 2013.

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

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

[16] 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," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

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

[18] P. J. Kenny, T. Stafylakis, P. Ouellet., Md. J. Alam, and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitraty Duration," in *IEEE International Conference on Audio, Speech and Signal Processing*, 2013.

[19] Y. Lei, M. McLaren, L. Ferrer, and N. Scheffer, "Simplified vts-based i-vector extraction in noise-robust seaker recognition," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[20] 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 *ISCA Interspeech*, 2013.

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

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

[23] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," in *ISCA Speaker Odyssey*, 2006.