Towards Duration Invariance of i-Vector-based Adaptive Score Normalization

Andreas Nautsch[*][†][‡], *Christian Rathgeb*[†], *Christoph Busch*[†],
Herbert Reininger[*] *and Klaus Kasper*[‡]

[*] atip Advanced Technologies for Information Processing GmbH, Frankfurt, Germany
{andreas.nautsch,herbert.reininger}@atip.de

[†] da/sec Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
[‡] Department of Computer Science, Hochschule Darmstadt, Germany
{christian.rathgeb,christoph.busch,klaus.kasper}@h-da.de

Abstract

It is generally conceded that duration variability has huge effects on the biometric performance of speaker recognition systems. State-of-the-art approaches, which employ i-vector representations, apply adaptive symmetric (AS) score-normalizations to improve the performance of the underlying system by using specific statistics on reference and probe templates obtained from additional datasets. The incorporation of duration information turns out to be vital in order to prevent a significant raise of entropy, since variation and likely a reduction of the signal duration from reference to probe samples is unpredictable.

In this paper we propose a duration-invariant extension of the AS-Norm, which is capable of computing more robust scores over a wide range of duration variabilities. The presented technique requires less computational effort at the time of speaker verification, and yields a 19% relative-gain in the minimum detection costs on the current NIST i-vector challenge database, compared to the provided NIST i-vector baseline system.

Keywords: biometrics, speaker recognition, i-vector, score normalization, duration invariance

1. Introduction

In past years speaker recognition has been incorporated in governmental, forensic, and industry applications [1] with a widespread scope ranging from court-cases [2] over preventing contact center frauds [3] to key security solutions for high-secure financial transactions [4]. Within conventional speaker recognition systems characteristic traits of an individual's voice are extracted in order to compare them against voice templates of known identities, *i.e.* speakers can either be verified or identified.

Recent studies demonstrated the feasibility of text- and language-independent speaker recognition by clustering the acoustical features space using *Gaussian Mixture Models* (GMMs), where the resulting universal cluster is referred to as *Universal Background Model* (UBM) [5, 6]. A speaker's feature space is then derived by a mean-only UBM adaptation with respect to the speaker's sample where the resulting mean-vector characterising a speaker's sample is defined as *supervector* [5, 6]. By analysing characteristic factors of the supervector offset from the UBM means, denoted by $\vec{\mu}_{\text{UBM}}$, Dehak *et al.* [7] introduced the *identity-vector* (i-vector) approach, which decomposes a speaker- and sample-dependent supervector $\vec{s}$ into a low-dimensional high-discriminative i-vector $\vec{i}$ by using a total variability matrix $\mathbf{T}$ which is trained by all prior-observed variational speaker and channel effects:

$$\vec{s} = \vec{\mu}_{\text{UBM}} + \mathbf{T}\vec{i}. \tag{1}$$

Consequently, i-vectors represent adequate features within a speaker-personalized space.

1.1. Motivation and Contribution

Presence of speech signal noise, which can occur due to *e.g.*, environmental noise, different microphones, channel-effects, within-speaker variabilities such as ageing, or duration-mismatches resulting in bad-estimated speaker subspaces, causes insufficiently estimated supervectors and i-vectors. In order to establish a robust speaker recognition systems increasing intra-class speaker variabilities need to be reduced towards a minimum.

This paper places emphasize on the reduction of i-vector noise arising due to duration variabilities. Effects of duration mismatches between enrollment and verification samples on i-vectors have been evaluated in past years pointing out that especially on short-term samples entropy rises much more than on long-term samples, which deliver sufficient statistics for i-vector extraction [8, 9].

Recently, i-vector performances have been analyzed with respect to sample durations and the according acoustical space [10]. A linear interrelation between the logarithmic duration and the amount of unique phone classes has been reported, *i.e.* the existence of so-called *acoustic holes* has been claimed depending on a sample's duration, which actually strongly influences the statistical sufficiency of estimating speaker subspaces. As a consequence, it has been suggested to evaluate score-calibration methods according to logarithmic duration classes.

Since there are different variations according to the duration classes, duration-based processing is very effective as we will show on the 2013–2014 NIST i-vector challenge where we applied a standard AS-norm to the NIST baseline system and extended the AS-norm by duration-sensitive development i-vector comparisons. By comparing i-vectors of the same duration-range, variations due to duration mismatches can be estimated and normalized more effectively.

1.2. Organization of Work

This paper is organized as follows: Sect. 2 summarizes relevant related work regarding duration mismatch compensation.

In Sect. 3 the proposed duration-based extension of the standard AS-norm will be presented in detail. Experimental results in terms of biometric performance and evidence strength are presented in Sect. 4. In Sect. 5 conclusions are drawn and future work is discussed.

2. Related Work

Karam *et al.* [11] analyzed cohorts, which are the nearest to one speaker for the purpose of performing AS-norm on the i-vector system suggested by Dehak *et al.* [7]. Mandasari *et al.* [8] evaluated i-vector systems using AS-norm[1] with respect to different sample durations. The authors demonstrated that basic i-vector systems significantly suffer from duration mismatches in terms of forensic applications. By employing the standard AS-norm, gains in evidence strength and performance could be obtained over several duration mismatch groups, by limiting evaluations to full-duration i-vectors. However, although gains were also yielded on short-duration samples, the vast majority of these systems tend to suffer from mis-calibration [8, 10, 12].

Kanagasundaram *et al.* [9] and Sarkar *et al.* [13] examined *Gaussian Probabilistic Linear Discriminant Analysis* (GPLDA) as a scoring alternative to the basic cosine comparator with focus on short-duration samples. GPLDA scores the likelihood of two i-vectors by a prior trained Gaussian model of i-vector between- and within-variances. Therefore, GPLDA assumes hidden speaker within- and between-variation factors $\vec{x}, \vec{y}_r$ for an extracted i-vector $\vec{i}$. Additional i-vector noise is compensated by these factors together with a-priori trained within- and between-variabilities $\mathbf{U}_r, \mathbf{V}$, such that a more robust i-vector representation can be obtained by [14]:

$$\vec{i} = \vec{\mu} + \mathbf{U}_r\vec{y}_r + \mathbf{V}\vec{y} + \vec{\epsilon}_r, \tag{2}$$

where $\vec{\mu}$ represents the development i-vector's mean, and $\vec{\epsilon}_r$ are standard Gaussian distributed residuals. A log-likelihood ratio (LLR) score is then obtained by estimating, whether the two i-vectors were emitted by same speaker or not, by assuming Gaussian distributed i-vectors. In order to compensate duration mismatches and variabilities as additional noise, GPLDA was additionally trained with low-durational samples in [9, 13, 14]. In general, more robust systems could be established, however, these systems yield huge performance losses with respect to lower durational probe samples.

Hasan *et al.* [10] analyzed effects of template and probe samples with respect to the acoustical feature space. They reported a linear dependency between the logarithmic duration and the amount of unique phone classes observed within a sample. Hence, they evaluated i-vector GPLDA performance with respect to duration groups, which were set up logarithmically. They improved the recognition robustness in terms of the actual detection cost by using score-calibration methods employing template and probe durations as quality measurements. Cumani *et al.* [15] performed a linear i-vector length normalization before GPLDA, which uses posterior distribution information of a-priori known i-vectors and accordingly projects i-vectors. Performance gains were reported on duration-variant scenarios.

Building on the approach in [10], Mandasari *et al.* [12] proposed more score-calibration methods taking template and probe durations d_t, d_p into account by using *Quality Model Functions* (QMFs) in order to reduce recognition entropy. For this purpose they trained calibration function parameters by linear regression, *i.e* the original score S is recalibrated to S',

$$S' = x_0 + x_1\, S + x_2\, \mathrm{QMF}(d_t, d_p), \tag{3}$$

where $x_{0,1,2}$ are parameters to be determined by linear regression using an additional database. Both, Hasan *et al.* [10] and Mandasari *et al.* [12], improved recognition robustness by reducing entropy employing score-calibration methods and the GPLDA scoring in order to compensate for noise.

Other researches emphasized on earlier processing stages: Fatima and Zheng [16], and Zhang *et al.* [17] propose phone-based speaker modeling by Gaussian-Mixture-Models which could be extended to phone-based i-vectors that would extend computational costs on signal processing compared to the standard i-vector approach. Stadelmann and Freisleben [18] discussed the usage of dimension-decoupled UBMs to reduce over-fitting of the acoustical space clustering. Hautamäki *et al.* [19] suggested minimax i-vector extractors to reduce mismatches within an i-vector neighbourhood. Since all these approaches are applied on processing stages before an i-vectors exists, they are not applicable towards the 2013–2014 NIST i-vector challenge, thus we emphasize later-stage noise reduction techniques.

However, if noise was produced by system processing, then score-calibration and exhaustive GPLDA training phases might deliver more significant gains in reducing error-propagation effects than in increasing i-vector performance abilities. In this paper we follow on a rather simple yet effective approach, extending the standard AS-norm by duration-invariant statistical treatments, which increase performance and omit entropy-emission.

3. System Architecture

The proposed system relies on (1) an i-vector baseline system on which (2) AS score-normalization is applied. In order to compensate for effects of varying durations after i-vector extraction the AS-norm will be applied in a (3) probe-duration-sensitive manner. Fig. 1 depicts the general system design, which will be described in detail in the following subsections.

3.1. i-Vector Baseline System

The i-vector baseline system is designed according to the NIST baseline system of the 2013–2014 i-vector challenge which takes benefits of recent methodologies in i-vector processing such as mean-subtraction, whitening transformation and length-normalization [20, 21, 22], i.e. i-vectors can be interpreted as unit-vectors.

The i-vector means $\vec{i}_{\mu_{\text{dev-set}}}$ represent an a-priori average offset of characteristic factors obtained from the UBM. By applying mean-subtraction the i-vector space is centered. However, i-vector elements, as the space axes, are correlated due to GMM-supervector element-correlations which occur due to the mean-concatenation of the GMM joint-mixtures. Hence, whitening is applied in order to transform correlated data into uncorrelated data exhibiting uniform variance, *i.e.* i-vectors are transformed to an uncorrelated space where the origin represents the average UBM-supervector deviation. Accordingly a whitening matrix $\mathbf{W}_{\text{dev-set}}$ is computed on a-priori known i-vectors of the development set (dev-set), such that an eigen-decomposition of the i-vector variances is used to transform the i-vector covariance matrix into an identity matrix.

[1] In the paper they refer to AS-norm as *normalized cosine kernel*.

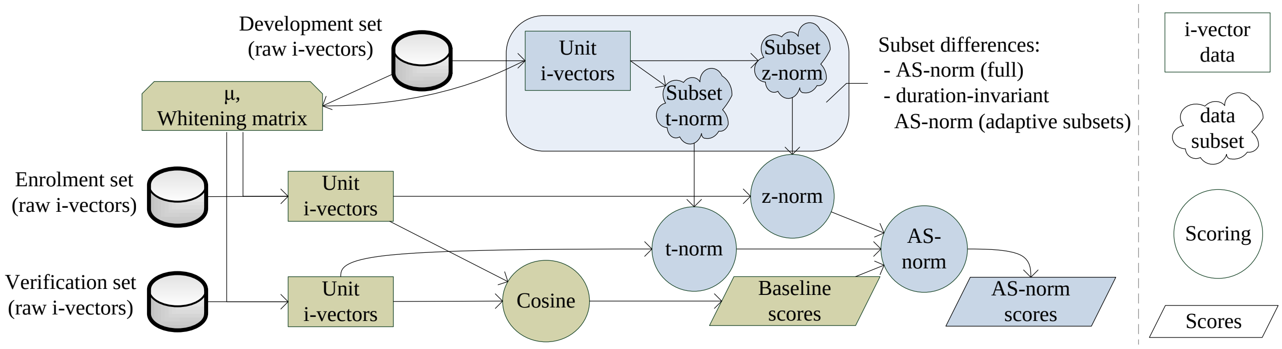


Figure 1: Basic operation mode of the proposed duration-sensitive speaker recognition system.

In order to deal with non-Gaussian behaviors, in the baseline system length-normalization is applied on the i-vectors as well [20], *i.e.* i-vectors can be also interpreted as features representing unit vectors in a speaker-characterizing space. Raw i-vectors $\vec{i}_{\text{raw}}$ are transformed into unit i-vectors $\vec{i}_{\text{unit}}$ applying the following equation:

$$\vec{i}_{\text{unit}} = \frac{(\vec{i}_{\text{raw}} - \vec{i}_{\mu_{\text{raw,dev-set}}})\mathbf{W}_{\text{raw,dev-set}}}{||(\vec{i}_{\text{raw}} - \vec{i}_{\mu_{\text{raw,dev-set}}})\mathbf{W}_{\text{raw,dev-set}}||} \tag{4}$$

where we will further denote $\vec{i} = \vec{i}_{\text{unit}}$ to ease notations.

Speaker references are created by averaging multiple enrollment i-vectors resulting in noise-robust templates [22, 23], which can be further interpreted as a sample-concatenated simulation where higher-sufficient Baum-Welch statistics are averaged, such that more speaker-characterizing i-vectors are extracted. At the time of verification the cosine similarity comparison between template and probe i-vectors $\vec{i}_t, \vec{i}_p$ is used according to the NIST baseline system [22]:

$$S(\vec{i}_t, \vec{i}_p) = \frac{\vec{i}_t^{\,T}\, \vec{i}_p}{||\vec{i}_t||\,||\vec{i}_p||} \tag{5}$$

where the i-vectors are already length-normalized, *i.e.* only the numerator term of Eq. 5 is required for score computations.

3.2. AS-Norm

For the purpose of applying standard score-normalization methods by preserving the symmetry between i-vectors, Kenny [24] introduced the *symmetric normalization* (s-norm). Thereby the *zero score-normalization* (z-norm) computes the score mean $\mu_{\text{z-norm}}$ and standard deviation $\sigma_{\text{z-norm}}$ of a template i-vector compared against an i-vector collection $\mathfrak{Z}$, and the *test score-normalization* (t-norm) compares similar parameters $\mu_{\text{t-norm}}, \sigma_{\text{t-norm}}$ of a probe i-vector against an i-vector collection $\mathfrak{T}$. Hence, a verification score S can be normalized by centering impostor scores having unit variance by known impostor score distributions with respect to a template i-vector and of a probe i-vector as if it was an impostor i-vector,

$$S' = \frac{1}{2}\left(\frac{S - \mu_{\text{z-norm}}}{\sigma_{\text{z-norm}}} + \frac{S - \mu_{\text{t-norm}}}{\sigma_{\text{t-norm}}}\right). \tag{6}$$

The AS-norm S' differs from s-norm by the scores which are used to compute the z/t-statistics: rather than using all scores, only the most competitive scores (*e.g.* top-100) are applied to model according speaker cohorts. Dehak *et al.* [25] applied the AS-norm on i-vectors and showed that the score normalization can be already applied on comparison-level as a normalized cosine scoring,

$$S(\vec{i}_t, \vec{i}_p) = \frac{(\vec{i}_t - \vec{i}_{\mu_{\text{z-norm}}})^T(\vec{i}_p - \vec{i}_{\mu_{\text{t-norm}}})}{||\mathbf{\Sigma}_{\text{z-norm}}\vec{i}_t||\,||\mathbf{\Sigma}_{\text{t-norm}}\vec{i}_p||} \tag{7}$$

where $\vec{i}_{\mu_{\text{z-norm}}}, \vec{i}_{\mu_{\text{t-norm}}}$ denote mean i-vectors of z- and t-norm sets, and $\mathbf{\Sigma}_{\text{z-norm}}, \mathbf{\Sigma}_{\text{t-norm}}$ are the square ”‘root of”’ of according diagonal covariance matrices.

3.3. Proposed Duration-invariant Approach

In order to build upon the idea of only taking significant comparisons into account, AS-norm is adapted to differentiate between probe sample durations. As previously mentioned, the presence of acoustic holes increases the entropy of shorter voice samples, which motivates the construction of different i-vector sufficiency-classes. Hence, the AS-norm is extended such that only comparisons are used for AS-parameter estimation that have the same quality as the current probe presented for verification.

In terms of duration as a quality metric, Q quality classes can be denoted as: $\mathfrak{Q} = \{\Lambda_0, \ldots, \Lambda_Q\}$ representing i-vector sufficiency classes. Samples are then associated by their logarithmic duration d_s to a sufficiency class Λ_c by the lowest log-duration distance,

$$\arg_{\Lambda_c} \min |\log(d_s) - \log(d_{\Lambda_c})|. \tag{8}$$

3.3.1. i-vector sufficiency classes

In the proposed system duration-based groups are defined for the sufficiency classes, where we limit the number of quality classes to $Q = 5$, *i.e.* obtained results can be directly compared to those reported in [10, 12]. It was found that evaluations carried out for the adaptive log-duration range from Eq. 8 yielded no significantly different results. Thus, sufficiency classes are denoted according to the researches on acoustic holes of Hasan *et al.* [10] and Mandasari *et al.* [12] and summarized in Table 1, where Λ_{full} is intended to comprise all expected high-sufficient i-vectors which might cause non-optimal results, but preserves low-computation efforts. This configuration is adequate for the purpose of verifying the method.

3.3.2. Parameter Estimation

For the z- and t-norm parameter AS-cohorts are pre-selected in various ways:

Table 1: Sufficiency classes and corresponding durations.

Sufficiency class	**Duration**
Λ_5	0–5 sec
Λ_{10}	5–10 sec
Λ_{20}	10–20 sec
Λ_{40}	20–40 sec
Λ_{full}	$\geq$ 40 sec

- z-norm simulates impostor verifications on averaged enrollment templates, thus only $\mathfrak{Z}$ i-vectors will be used which are belong to the same sufficiency class as the probe i-vector:

$$\mathfrak{Z} = \{\vec{i}_{\Lambda_{d_p}} \,|\, \max_{\text{top100}} S(\vec{i}_t, \vec{i}_{\Lambda_{d_p}})\}, \tag{9}$$

- t-norm simulates impostor verifications comparing the probe i-vector to templates of the development set, where enrolled speakers have full i-vectors, where the vast majority of durations are higher than 60 seconds, *i.e.* only $\mathfrak{T}$ i-vectors will be used extracted from samples with longest durations:

$$\mathfrak{T} = \{\vec{i}_{\Lambda_{>60}} \,|\, \max_{\text{top100}} S(\vec{i}_t, \vec{i}_{\Lambda_{>60}})\}. \tag{10}$$

3.3.3. Score estimation and expected improvements

The proposed duration-adaptive extension of AS-norm normalizes the scores according to Eq. 6. By placing emphasis on duration-based sufficiency classes, recognitions are treated duration-invariant, *i.e.* normalized scores are expected to be distributed without creating entropy due to duration-mismatches. Further, an overall improvement can be expected, since scores of all sufficiency classes are normalized to more similar distributions of genuine and impostor scores. As a consequence, no additional entropy is expected,which could arise due to score-distribution mismatches by fixed across-classes thresholds.

Fig. 2 illustrates how duration-differing samples will be processed by the proposed AS-norm extension.

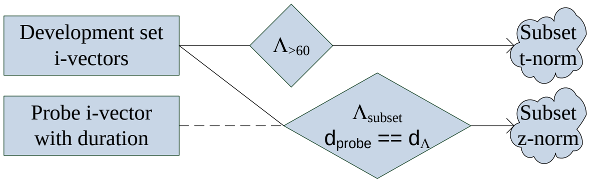


Figure 2: Processing duration-differing samples by suggested duration-based AS-norm extension.

4. Experimental Evaluation

Experiments are carried out on the 2013–2014 NIST i-vector challenge dataset [22] in order to evaluate the baseline, the standard AS-norm, and the duration-based AS-norm extension. We performed ten 5-fold cross-validations[2] on the enrollment database, and we submitted each system also on the i-vector challenge where preliminary results were computed by NIST using 40% of the whole evaluation set. The data sets were only used to evaluate the method and not to tune the submitted system.

4.1. Experimental Set-up

The NIST i-vector challenge dataset consists of $1\,306$ speaker identities within enrollment and verification sets. For each identity 5 enrollment i-vectors are given with the according sample duration. The verification set contains $9\,634$ probe i-vectors with the according sample duration as well. Further, a development set of $36\,572$ independent i-vector with sample durations is given for feature space estimations, independent of the evaluation data[3].

Focusing on performance evaluation, we place emphasize on the biometric recognition performance in terms of the Equal-Error-Rate (EER), and the false non-match rate at a 1% false match rate (FMR100). In accordance to the ISO/IEC IS 19795-1 [26] the FNMR of a biometric system defines the proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample. By analogy, the FMR defines the proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template. As score distributions overlap EERs are obtained, i.e. the system error rate where FNMR = FMR. Further, we estimate the entropy and biometric performance in terms of the application-dependent[4] minimum detection cost function [22]

$$\text{minDCF} = \min\, \text{FNMR} + 100\,\text{FMR}, \tag{11}$$

and the application-independent entropy by the log-likelihood ratio cost [28] of genuine and impostor scores SG, SI

$$C_{\text{llr}} = \frac{\sum_{g \in SG} \text{ld}(1 + \frac{1}{e^g})}{2|SG|} + \frac{\sum_{i \in SI} \text{ld}(1 + e^i)}{2|SI|}. \tag{12}$$

In accordance to ISO/IEC IS 19795-1 [26], we refer to FMR and FNMR instead of FRR and FAR, since we are evaluating the algorithmic performance without knowing the actual failure to enroll (FTE) and failure to capture (FTC) rates, which are effecting the biometric system performance rates FRR and FAR, respectively.

4.2. Data Analysis

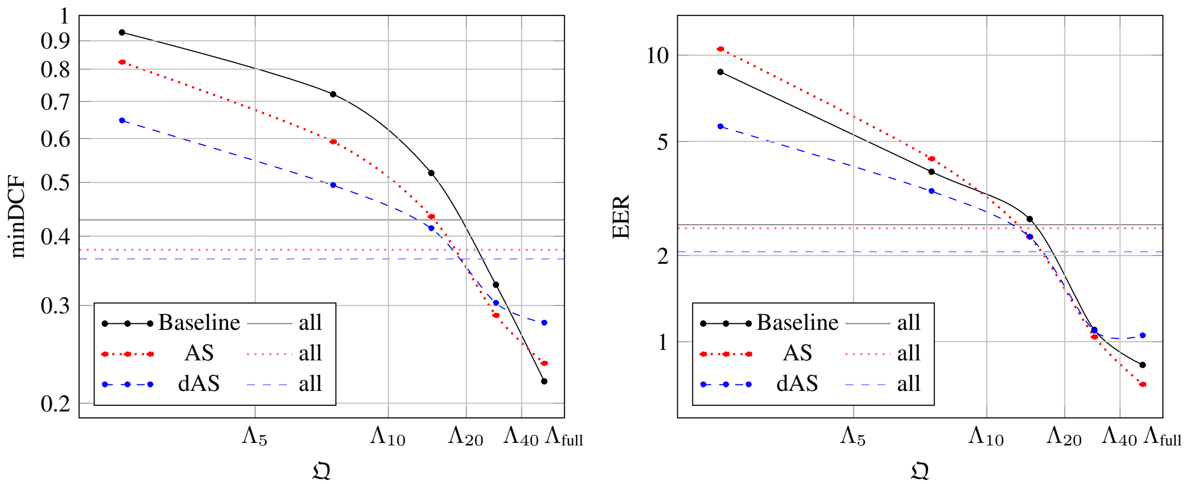
The provided i-vectors exhibit 600 dimensions, and their according sample durations are log-normal distributed as shown in Fig. 3. Most of the development sample durations are in the 20–40 second range, *i.e.* these samples are influencing development set based i-vector processing such as mean-subtraction and whitening.

The vast majority of development samples are located in Λ_{full} (34.7%), Λ_{40} (31.1%), and Λ_{20} (23.1%), then: Λ_{10} (9.0%), and Λ_5 (2.1%). Intentional, all i-vectors have been centralized to the origin by mean-subtraction in the preparation of the baseline system, but an unpaired Student t-test of independence showed that i-vector elements have significantly different mean-values compared between all development set i-vectors and with respect to each sufficiency class. Table 2 compares the

[2] On each validation run one enrollment i-vector was randomly taken as a probe while the remaining i-vectors were used to create a template.

[3] The usage of information about other trials within the evaluation data is not allowed by the NIST challenge protocol [22].

[4] NIST set the i-vector challenge operating point similar to NIST SRE'10 at an effective prior $\tilde{\pi} = \frac{1}{101}$ [22, 27] with a Bayes threshold of $\eta \approx 4.6$.



(a) MinDCF

(b) EER

Figure 4: Biometric performance of i-vector sufficiency classes: (a) minDCF, (b) EER.

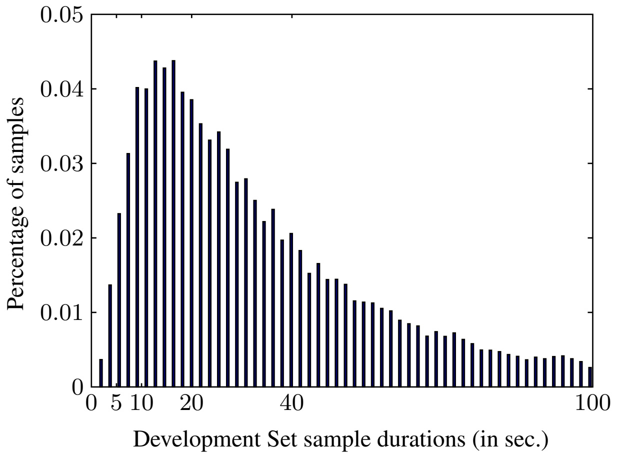


Figure 3: Log-normal distributed sample durations of development set data with respect to i-vector sufficiency classes.

amount of significantly independent i-vector elements according to their sample durations assuming equal variance[5].

Table 2: Student t-test of independent i-vector elements with respect to sufficiency classes.

dev-set	all	Λ_{full}	Λ_{40}	Λ_{20}	Λ_{10}
Λ_5	84	141	91	66	44
Λ_{10}	142	230	140	70	
Λ_{20}	132	246	118		
Λ_{40}	35	180			
Λ_{full}	172				

Once mean-subtraction and whitening has been applied, Λ_{40} i-vectors exhibit the lowest significant offset to the space origin by having the second most impact on both i-vector processing due to their representative amount. Further, the most-sufficient i-vectors have the greatest gap compared to all development set i-vectors and to each sufficiency class with at least $140/600$ significant different mean positions. Hence, within the subspace of Λ_{40} i-vectors seem to be between the subspaces of high-insufficient and high-sufficient i-vectors. An opposite effect could be observed on short-duration samples, where the according i-vectors have larger mean-differences to i-vectors of more than 20 seconds than to i-vectors of comparable short duration samples (less than 20 seconds). This effect may be caused due to high variability of insufficient estimated i-vectors of short-duration samples, *i.e.* i-vectors of less than 20 second samples are distributed in subspaces that are more close to themselves than to more-sufficiently estimated i-vectors.

That is, offset vectors can be assumed for each sufficiency group, which effect the cosine score values due to angle changes between i-vectors[6]. These facts underline the need for compensating scoring statistics with respect to sample durations.

4.3. Performance Evaluation

Focusing on the baseline system the highest performance loss in terms of minDCF is observed for low-durational samples, see Table 3. As it can be seen, Λ_5 i-vectors yielded the most expensive detection costs with 0.932 which is very close to a random recognizers performance of $minDCF = 1$. I-vectors stemming from the class with the longest sample duration yielded the best observed minDCF, *i.e.* 0.219. However, on all other quality classes of insufficient i-vectors both AS-normalizations yield significant gains where the duration-invariant AS-norm outperforms both other systems on samples shorter than 20 seconds. On 20–40 second samples both normalizations could outperform the baseline approach, where AS-norm without duration-sensitive extension achieved the best minDCF for Λ_{40} i-vectors. Hence, AS-norm is necessary on insufficiently estimated i-vectors, and the proposed duration-based extension can yield up to 19.1% more relative-gain than the standard AS-norm.

In terms of biometric recognition performance both AS-

[5] Results of an unpaired Student t-test assuming un-equal variances yielded negligible differences in the results.

[6] Which actually is additive noise that should be well-compensable by, *e.g.* GPLDA scoring.

Table 3: Duration group performances: avg. minDCF.

System	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	0.932	0.721	0.520	0.327	**0.219**
AS-norm	0.824	0.592	0.434	**0.288**	0.236
dAS-norm	**0.646**	**0.494**	**0.413**	0.303	0.279

norm approaches outperform the baseline as well, see Table. 4. Again, the proposed duration-invariant AS-norm yields significant gains on samples shorter than 20 seconds on which a performance break-down for the standard AS-norm can be observed. However, on higher-sufficient i-vectors the standard AS-norm outperforms both other systems, which could be caused due to the non-duration-invariance within the Λ_{full} i-vectors. EER and minDCF performance comparisons among quality classes $\mathfrak{Q}$ are shown in Fig. 4.

Table 4: Duration group performances: avg. EER.

System	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	8.74	3.92	2.68	1.10	0.83
AS-norm	10.51	4.35	**2.32**	**1.04**	**0.71**
dAS-norm	**5.63**	**3.35**	**2.32**	1.09	1.05

Across the entire set of classes the proposed duration-based AS-norm outperforms both other systems, see Table 5. In summary, the proposed duration-invariant AS-norm yields a 19.5% relative-gain in EER, a 32.6% relative-gain in FMR100, and a 15.0% relative-gain in minDCF compared to the baseline system on the cross-validation. Further, the duration-invariant AS-norm significantly outperforms the standard AS-norm which can also be seen in Fig. 5.

Table 5: System performances: avg. EER, FMR100, minDCF.

System	EER	FMR100	minDCF	Challenge[7]
Baseline	2.56	5.15	0.428	0.386
AS-norm	2.49	4.48	0.378	0.331
dAS-norm	**2.06**	**3.47**	**0.364**	**0.312**

The results were approved by the preliminary evaluation of the 2013–2014 NIST i-vector challenge, where the application of the standard AS-norm resulted in a 14.2% relative-gain, and the duration-invariant extension resulted in a 19.2% relative-gain in minDCF.

Fig. 5 compares the best cross-validation systems according to minDCF within a Detection Error Trade-off diagram. The duration-invariant AS-norm improves the biometric performance of the baseline system at all operating points, while the standard AS-norm mainly yields gains in high-secure regions, *i.e.* operating points at low FMRs. In this regions both AS-normalizations exhibit equal recognition accuracy.

Hence, the proposed duration-invariant AS-norm extension is applicable to a larger range of scenarios compared to the standard AS-norm. While the duration-invariant AS-norm only obtains slightly lower error-rates on minDCF-operating points compared to the standard AS-norm, another advantage of the duration-invariant treatment is observed within entropy evaluations.

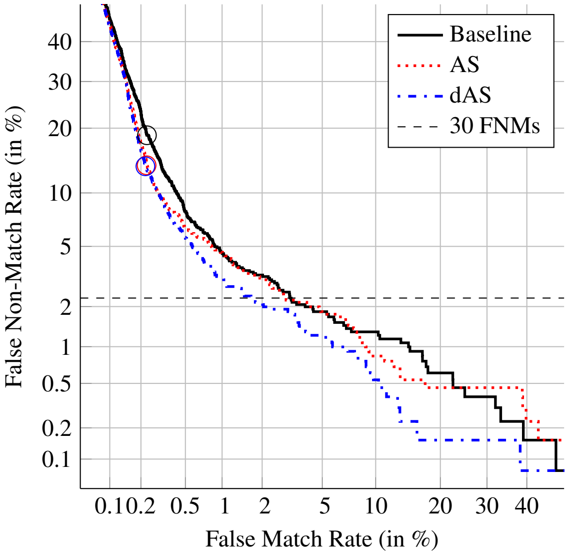


Figure 5: Systems detection error tradeoff: best systems from 10 cross-validations according to their minDCF.

4.4. Entropy Evaluation

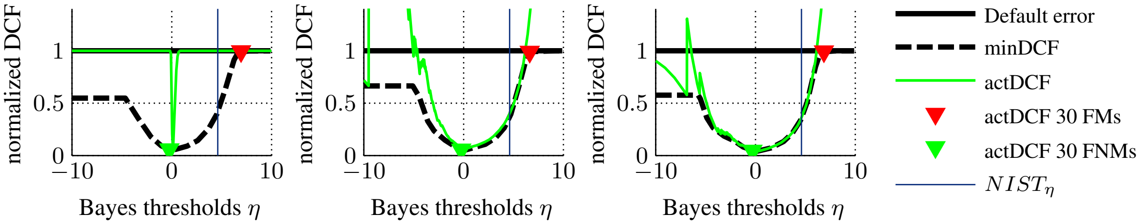
Table 6 compares the total C_{llr} of the three systems over all scores, and among each quality class. On Λ_5 i-vectors the baseline and the standard AS-norm perform similar to or worse than a random recognizer, and on samples having more than 5 seconds the standard AS-norm significantly outperforms the baseline system. On high-sufficient i-vector the lowest application-independent entropy was measured for the standard AS-norm with $C_{\text{llr}} = 0.05$, representing a very low cost of the LLR-scores. However, on sample durations lower than 40 seconds the duration-invariant AS-norm outperforms both other approaches by yielding a maximum LLR cost of $C_{\text{llr}} = 0.35$ on high-insufficient Λ_5 i-vectors. Overall the suggested AS-norm extension exhibits the lowest application-independent entropy by yielding relative-gains of 88.8% and 41.2%, respectively.

Table 6: Average entropy comparison: all scores & duration-groups.

System	all	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	0.89	0.95	0.93	0.92	0.89	0.86
AS-norm	0.17	1.18	0.41	0.18	0.08	**0.05**
dAS-norm	**0.10**	**0.35**	**0.20**	**0.11**	**0.07**	0.07

Fig. 6 illustrates the C_{llr} gains on normalized DCFs or likewise normalized Bayesian entropy plots, where the actual DCF (actDCF) represents application-dependent entropy, and the minDCF represents application-dependent entropy on a well-calibrated system — in these terms C_{llr} represents the area under actDCF, since we want to place emphasize on robustness, *i.e.* systems which do not require score-calibration. Due to the cosine scoring most scores of the baseline system lie within the range $[-1, +1]$, hence, the lowest DCF. The smallest difference between actual and minimum DCF was observed on

[7] The results were obtained by the 2013–2014 NIST i-vector online leaderboard which comprised 40% of the total evaluation data.



(a) Baseline (b) AS-norm (c) dAS-norm

Figure 6: Entropy comparison of (a) the baseline system, (b) the standard AS-norm and (c) the proposed duration-invariant AS-norm.

$\eta \approx 0$, on any other operating point the baseline system is effected by huge mis-calibrations. Calibration-improvements were gained by the standard AS-norm, which delivers adequate calibration for a different application-points (actDCF curve being equal to minDCF curve). However, the suggested duration-invariant score-normalization yields well-calibrated scores on the vast majority of application-points, which have significant error-rates. That is, the proposed duration-invariant enables an enhanced statistical treatment of quality classes, which is approved by a very low overall entropy emission in terms of C_{llr}.

4.5. Discussion

Quality classes of i-vector sufficiency were motivated by assuming acoustical holes depending on the logarithmic sample duration. By observing i-vector mean offsets between the quality classes $\mathfrak{Q}$, the need of duration-invariant recognition processes were empirically motivated in order to compensate for i-vector subspace mismatches. Hence, the AS-norm was extended with respect to duration-based quality classes as proposed in Sect. 3.3.

The experimental results showed that statistical effects of acoustical holes causing entropy are easy-compensable by analyzing their i-vector subspace variations according to same-shaped quality classes. Hence, additional processing-entropy was prevented for many operating points, and significant performance gains were yielded on short-duration samples as well. However, on high-sufficient i-vectors the standard AS-norm provides slightly better results, thus combined systems are considered promising with respect to recognition performance. Furthermore, more detailed separation of quality classes within Λ_{full} are expected to yield further gains within the proposed duration-invariant AS-norm.

Placing emphasize on the computational complexity the standard AS-norm requires all 36 572 development set i-vectors for either of the z-norm and t-norm sets in order to determine the top100 cohorts. In contrast, the duration-invariant extension utilizes at most 34.7% of the data amount for z-norm for Λ_{full} quality class normalizations, and 19.4% of the complete development set. Thus, proposed extension turns out to be highly suitable to units having less computational resources.

5. Conclusion and Future Work

The proposed duration-invariant extension of AS-norm is shown to exhibit high performance by robust evidence strength, hence we assume the method to be suitable for industry and forensic application use cases. Entropy in short-term duration classes could be reduced significantly, and a 19% relative-gain in biometric performance can be observed compared to the baseline system, proving the soundness of the presented approach. Further, the overall forensic evidence strength could be significantly increased, reducing the actual LLR cost to $C_{\text{llr}} = 0.10$. The dAS-norm applies additional information of verification attempts with known speaker identities that are very similar according to the used comparator, *e.g.* by the cosine similarity score, such that comparator-based entropy can be significantly reduced, *e.g. measured by the minDCF metric*: the i-vectors are analyzed w.r.t. similar conditioned i-vectors subspaces, rather than to i-vector collections of various extraction sufficiency.

Building upon our reproducible technique, future research might investigate template and probe normalizations as performed by Eq. 7 where quality-class-dependent mean offsets are vanished to achieve higher recognition performances. Though, limitations are expected in terms of the i-vector subspace coverage by the dAS-norm i-vector collection and in terms of the comparator's performance potential which might be outperformed by considering more fitted speaker comparison techniques. Hence, more sufficient comparators such as GPLDA or the two-covariance model [29], GPLDA's dot-product variation for fast scoring, can be applied on low-entropy i-vectors to concern more signal-based rather than processing-based entropy.

Further, duration-based quality classes can be investigated on invariant treatments towards their specific characteristics as i-vector subspaces on earlier processing stages such as duration-based i-vector extraction techniques. The proposed method should be easy transferable to other scoring techniques, such as the GMM-UBM or GPLDA comparators.

6. Acknowledgment

This work has been partially funded by the Center for Advanced Security Research Darmstadt (CASED).

We would like to thank the database creators from the National Institute of Standards and Technology (NIST), the Linguistic Data Consortium (LDC), the Massachusetts Institute of Technology: Lincoln Laboratory (MIT-LL), and the Johns Hopkins University (JHU), and NIST and the 2014 Speaker Odyssey committee for hosting and database distribution.

7. References

[1] A. K. Jain, A. Ross, and S.Prabhakar, "An Introduction to Biometric Recognition," in *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 2004.

[2] J. A. Batchelor, D. M. Lee, D. P. Banks, D. A. Crosby, K. W. Moore, S. H. Kuhn, T. Rodriguez, and A. B. Stephens, "Ivestigative report," Florida Department of Law Enforcement, 2012, Laboratory report on US law case: State of Florida v. George Zimmerman.

[3] D. Averbouch and J. Kahn, “Fraud Targets the Contact Center: What Now?,” Speech Technology Magazine, November 2013, White paper of NICE systems.

[4] SESTEK, “The Rise of Voice Biometrics as a Key Security Solution,” Speech Technology Magazine, July 2013, White paper of SESTEK.

[5] T. Kinnunen and H. Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors,” in *EURASIP Speech Communication*, 2010.

[6] A. Fazel and S. Chakrabartty, “An Overview of Statistical Pattern Recognition Techniques for Speaker Verification,” in *IEEE Circuits and Systems Magazine*, 2011.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2011.

[8] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “Evaluation of i-vector Speaker Recognition Systems for Forensic Applications,” in *ISCA Interspeech*, 2011.

[9] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “i-vector Based Speaker Recognition on Short Utterances,” in *ISCA Interspeech*, 2011, pp. 2341–2344.

[10] 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 (ICASSP)*, 2013.

[11] Z. N. Karam, W. M. Champbell, and N. Dehak, “Towards reduced false-alarms using cohorts,” in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2011.

[12] 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.

[13] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, “Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification,” in *ISCA Interspeech*, 2012.

[14] P. Kenny, T. Stafylakis, P. Ouellet, Md. J. Alam, and P. Dumouchel, “PLDA for speaker verification with utterances of arbitrary duration,” in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2013.

[15] S. Cumani, O. Plchot, and P. Laface, “Probabilistic linear discriminant analysis of i-vector posterior distributions,” in *IEEE International Conference on Audio, Speech and Signal Processing*, 2013.

[16] N. Fatima and T. F. Zheng, “Short Utterance Speaker Recognition — A research Agenda,” in *IEEE International Conference on Systems and Informatics (ICSAI)*, 2012.

[17] C. Zhang, X. Wu, T. F. Zheng, L. Wang, and C. Yin, “A K-Phoneme-Class based Multi-Model Method for Short Utterance Speaker Recognition,” in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012.

[18] T. Stadelmann and B. Freisleben, “Dimension-Decoupled Gaussian Mixture Model for Short Utterance Speaker Recognition,” in *IEEE International Conference on Pattern Recognition (ICPR)*, 2010.

[19] V. Hautamäki, Y.-C. Cheng, P. Rajan, and C.-H. Lee, “Minimax i-vector extractor for short duration speaker verification,” in *ISCA Interspeech*, 2013.

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

[21] 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 (CIARP)*, 2013.

[22] C. Greenberg, “The 2013-2014 Speaker Recognition i-vector Machine Learning Challenge,” 2013, NIST.

[23] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, “A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation,” in *ISCA Interspeech*, 2013.

[24] P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors,” in *ISCA Odyssey*, 2010.

[25] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, “Cosine Similarity Scoring without Score Normalization Techniques,” in *ISCA Odyssey*, 2010.

[26] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*, International Organization for Standardization and International Electrotechnical Committee, Mar. 2006.

[27] N. Brümmer and E. de Villers, “The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF,” 2011.

[28] N. Brümmer and J. du Preez, “Application-Independent Evaluation of Speaker Detection,” in *ISCA Odyssey: The Speaker and Language Recognition Workshop*, 2006.

[29] S. Cumani, N. Brümmer, L. Burget, and P. Laface, “Fast discriminative speaker verification in the i-vector space,” in *IEEE International Conference on Audio, Speech and*