Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Robustness of Quality-based Score Calibration of Speaker Recognition Systems with respect to low-SNR and short-duration conditions

Andreas Nautsch* Rahim Saeidi† Christian Rathgeb* 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

Degraded signal quality and incomplete voice probes have severe effects on the performance of a speaker recognition system. Unified audio characteristics (UACs) have been proposed to quantify multi-condition signal degradation effects into posterior probabilities of quality classes. In previous work, we showed that UAC-based quality vectors (q-vectors) are efficient at the score-normalization stage. Hence, we motivate q-vector based calibration by using functions of quality estimates (FQEs). In this work, we examine the robustness of calibration approaches to low-SNR and short-duration conditions utilizing measured and estimated quality indicators. Thereby, comparisons are drawn to quality measure functions (QMFs) employing oracle SNRs and sample duration.

In the robustness study, low-SNR and short-duration conditions are excluded from calibration training. The present analysis provides insights on the behavior of calibration schemes in combined conditions of high signal degradation and short segment duration regarding accurate approximation of idealized calibration. We seek calibration methods in order to parsimonious preserve robustness against unseen data.

A separate analysis is provided on duration- and noise-only scenarios as well as on combined duration and noise scenarios. QMFs and FQE reduce C_{mc} costs down to 5 – 6% of conventional calibration schemes if all conditions are known, and to 10 – 12% in the presence of unseen conditions.

Keywords: calibration, function of quality estimates, quality measure functions, robustness, multi-condition robustness

1. Introduction

In the vast majority of commercial and forensic application scenarios, robust handling of real world data is still a challenging topic. While commercial applications can place focus on known environments and permit sample re-acquisitions from the biometric data subject, the environmental set-up in forensic scenarios changes case by case. Furthermore, commercial applications need to cope more and more with varying conditions due to the rising demand for highly mobile applications facing unconstrained environmental conditions regarding sample acquisition processes.

Lately, the robustness of state-of-the-art speaker recognition systems was addressed regarding feature extractors [1, 2, 3, 4], uncertainty-aware comparators [5, 6], and score normalization and calibration schemes employing quality metrics [7, 8, 9]. In this paper, we study the calibration stage of a state-of-the-art speaker recognition system employing probabilistic linear discriminant analysis (PLDA) [10] of i-vectors [11]. Unified audio characteristic [12] based quality vectors (q-vectors) are estimated during i-vector extraction [8]. This work extends investigations conducted in [8] with respect to the following aspects: (1) the effectiveness of q-vector based calibration schemes, and (2) limitations regarding unseen conditions are examined by excluding conditions of low-SNR or short-duration from calibration training. Contrary to [8], where q-vectors were utilized for score normalization purposes, the focus of this work is put on score calibration schemes excluding score normalization. Thereby, we seek lower miscalibration costs compared to conventional calibration, which trains calibration functions on data stemming from an optimal condition (long duration, noise-free). This mismatched calibration scheme is expected to model low-SNR and short-duration conditions insufficiently, and thus to state the lower performance bound in our robustness analysis. In contrast, matched calibration is considered as optimal calibration, since calibration parameters are trained solely depending on each condition, respectively, which requires calibration functions to be trained for each condition. Hence, the complexity in terms of degrees-of-freedom increases on condition-matched calibration, when more conditions are considered and further, unseen conditions cannot be calibrated well. Therefore, *quality-based* calibration promises an adequate trade-off between model complexity and accurate approximation of condition-matched calibration in scenarios facing a wide range of combined noise and duration conditions.

This paper is organized as follows: Sec. 2 depicts the state-of-the-art and related work on quality-based calibration, and on those FQEs for calibration of recognition scores. A comprehensive analysis of calibration robustness regarding duration, noise and combined effects is provided in Sec. 3, and conclusions are drawn in Sec. 4.

2. Related Work

The focus of this work is put on score calibration of speaker recognition systems, thus a brief overview on the state-of-the-art is provided before related work on linear calibration and quality functions is depicted.

2.1. State-of-the-Art Speaker Recognition

Recent speaker recognition approaches rely on i-vectors, representing the characteristic speaker offset from an Universal Background Model (UBM), which models the distribution of acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCCs) [13]. Thereby, UBM components' mean vectors are

 doi: 10.21437/Odyssey.2016-52

Table 1: *Label scheme for combined duration and noise conditions, cf. [8].*

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

concatenated to a *supervector* $\vec{\mu}_{\text{UBM}}$. Speaker supervectors $\vec{s}$ are decomposed by a total variability matrix $\mathbf{T}$ into a lower-dimensional and higher-discriminant i-vector $\vec{i}$ as an offset to the UBM supervector $\vec{\mu}_{\text{UBM}}$:

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

The total variability matrix is trained on a development set using an expectation maximization algorithm [11, 14]. Then, i-vectors are projected onto a spherical space by whitening transform and length-normalization [10, 15]. State-of-the-art i-vector comparators belong to the Probabilistic Linear Discriminant Analysis (PLDA) family [15, 16]. PLDA comparators conduct a likelihood ratio scoring comparing the probabilities of the hypotheses (a) reference and probe i-vectors $\vec{i}_{\text{ref}}, \vec{i}_{\text{prb}}$ stem from the same source, or (b) stem from different sources. Therefore, speaker within and between variabilities are examined.

2.2. Different Environmental Conditions

Variations in signal quality, i.e. in the probe sample condition, result in different score distributions per condition [7, 17]. While systems are usually calibrated for known scenarios and in fix-condition environments, handling unconstrained conditions imposes well-calibrated decision thresholds among known and unseen conditions.

In this paper, we examine the 55 duration and noise conditions presented in [8]. In [8], SNR conditions stem from two noise sources: air conditioner (AC) and crowd (CROWD) noise. By degenerating voice samples from the I4U file list [18], combined signal degradation and observation incompleteness (short probe segment duration) effects are simulated, which are expected to represent the most common conditions, cf. Tab. 1.

2.3. Linear Calibration

Calibration maps raw recognition scores into the domain of well-calibrated scores. A linear calibration function adjusts the bias of scores S by an offset w_0 and scales S by a weight w_1, such that calibrated scores S' are computed as:

$$S' = w_0 + w_1\, S, \tag{2}$$

where w_0, w_1 are estimated by logistic regression [19, 20]. While non-linear approaches are able to calibrate well over wide application prior ranges, linear strategies may achieve well-calibration on more limited prior ranges [19]. However, in this work linear strategies are preferred, since they are expected to be more flexible regarding unseen data [19].

2.4. Quality Measure Functions

In scenarios targeting different conditions, the conventional linear calibration is observed to be prone to mis-calibration, when dealing with recognition scores originating from low-quality probe segments [7, 17]. Training calibration parameters condition-dependently leads to inconvenient effects, such as higher system complexity in terms of parameters to train. Quality measure functions (QMFs) [7, 21] were introduced in order to account for the quality of reference and probe samples in the score calibration process. In [7, 21], the QMFs are formulated as additional components in the linear calibration strategy. A QMF Q depends on reference (ref) and probe (prb) sample quality measures $\lambda_{ref}, \lambda_{prb}$:

$$S'_Q = w_0 + w_1\, S + \mathrm{Q}(\lambda_{ref}, \lambda_{prb}). \tag{3}$$

The QMF calibration can also be interpreted as a linear system fusion of the biometric comparison with a sub-system quantizing the quality of reference and probe.

Previous works [7, 17] introduced the following duration- and SNR-dependent QMFs:

$$Q_1 := w_2\, \log(d_{prb}), \tag{4}$$

$$Q_2 := w_2\, \mathrm{SNR}_{prb}, \tag{5}$$

$$Q_{1+2} := w_2\, \log(d_{prb}) + w_3\, \mathrm{SNR}_{prb}, \tag{6}$$

where $\mathrm{d}_{prb}, \mathrm{SNR}_{prb}$ denote the duration and SNR of the probe sample, respectively. Reference samples are assumed to stem from the clean/full (noise free long utterance) condition.

2.5. Function of Quality Estimates

Since measuring certain quality metrics such as SNR is difficult in low-quality conditions, alternative calibration schemes can rely on function of quality estimates (FQEs). The *quality estimators* can be obtained by using condition-dependent modeling of i-vectors.

2.5.1. Estimation of Unified Audio Quality Vectors

For the purpose of estimating quality in speaker recognition, unified audio characteristics [12] are utilized. Single multivariate Gaussian models $\Lambda_j \sim \mathcal{N}(\mu_j, \Sigma), j = 1, \ldots, 55$ are trained in original i-vector space for each quality condition as outlined in Tab. 1. The models have condition-dependent mean vectors μ_j 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 $\vec{i}$ represent a condition quality vector (q-vector) $\vec{q}$ [12], with entries:

$$q(j) = \frac{P(\vec{i}\,|\,\Lambda_j)}{\sum_{j=1}^{55} P(\vec{i}\,|\,\Lambda_j)}. \tag{7}$$

All i-vectors (ref, prb) are extended to a pair of an i-vector and a corresponding q-vector. Fig. 1 depicts average cosine score between q-vectors stemming from different conditions in the calibration training set. Same-conditions comparisons lead to comparably higher scores ($\simeq 1$) than cross-condition comparisons ($\simeq 0$). However, comparisons of alike-conditions rising from the same noise-type or SNR-level conditions cannot be easily accomplished, using a Gaussian model in i-vector space.

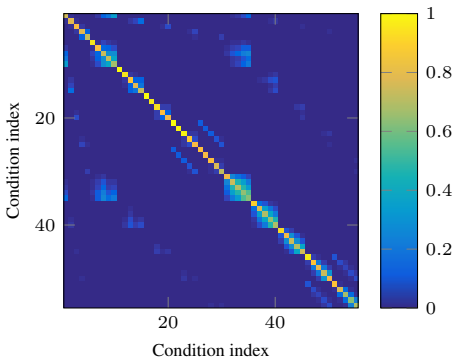


Figure 1: *Cross-condition comparison of q-vectors using an average cosine distance.*

2.5.2. Calibration based on q-vectors

Concerning multiple sample conditions, q-vector based calibration can be global (condition-pooled training), metadata based (pre-selection of matched conditions), and trial-based [9], i.e. samples employed during the calibration training are selected based on q-vector similarity to the probe sample's q-vector. Complementary to [9], we put also focus on two other conventional variants of global calibration schemes, in particular the condition-mismatched scheme (knowing only full/clean samples) and the condition-matched scheme, in which we assume an oracle knowledge of the matching condition is available.

In [9], trial-based calibration is proposed for examining a set of 14 distinct conditions comprising cross-language, cross-channel, noisy and reverberant effects. Where [9] aims at a wide range of condition types, distinct condition types are examined in this work in-depth. By selecting the 1000 most-alike samples for calibration training for each comparison, trial-based calibration causes a tremendous high degree-of-freedom on large-scale operations. Contrary, we propose an FQE-based calibration method in order to reduce the amount of total parameters involved in the calibration phase.

In [12], the proposed q-vector calibration scheme utilizes a symmetric, bilinear combination matrix $\mathbf{W}$ for conducting a quality similarity score between reference and probe q-vectors $\vec{q}_{\text{ref}}, \vec{q}_{\text{prb}}$:

$$Q_{\text{UAC}} := w_2 \, \vec{q}_{\text{ref}}^{\,T} \, \mathbf{W} \, \vec{q}_{\text{prb}}. \tag{8}$$

Where the estimation of $\mathbf{W}$ induces tremendously high degrees-of-freedom to the calibration stage, i.e. $\|\vec{q}\|^2$ parameters need to be estimated additionally: 3 025 $\mathbf{W}$ elements in order to cope with all of the 55 examined conditions. Low-rank estimates can be obtained by probabilistic principal component analysis. Following the QMF intention of parsimonious robustness against unseen conditions during calibration training, Q_{UAC} turns unfavorable by increasing condition amounts.

Based on the q-vector design, conditions can be identified by using q-vectors [8, 9, 12]. We propose the usage of the cosine distance between reference and probe q-vectors $\vec{q}_{\text{ref}}, \vec{q}_{\text{prb}}$ for score calibration as the FQE Q_{qvec}:

$$Q_{\text{qvec}} := w_2 \, \cos(\vec{q}_{\text{ref}}, \vec{q}_{\text{prb}}). \tag{9}$$

Compared to the calibration model proposed in [12], the proposed FQE requires far fewer calibration parameter estimations, i.e. one parameter w_2. In [12], covariance-alike matrices are required, thus even if diagonalized matrices are conducted, the degree-of-freedom equals the number of conditions leading in this setup to 55 additional calibration parameters to w_2. Which is not favorable for the purpose of facing robustness towards unknown conditions.

3. Analysis on Calibration Robustness

The scores for our experiments are derived from the baseline recognition system described in our recent work [8]. After examining baseline calibration results on the I4U evaluation subset, i.e. QMFs and FQEs are compared to mismatched and matched calibration schemes, we provide a study on the robustness of QMF and FQE calibration schemes against unseen conditions. In order to examine various performance effects of the calibration schemes, analyses are grouped by:

- variable duration: only conditions of 5 s, ..., full without noise are considered during calibration training and evaluation,
- variable SNR: only noisy conditions with full duration (including full/clean) are considered during calibration training and evaluation, and
- combined conditions: all 55 conditions are considered during calibration training and evaluation.

Since similar trends on CROWD and AC noise were found, comparable to [8], experimental results are only reported with respect to CROWD noise. In order to provide a compact overview, we also examine pooled conditions, i.e. the performance of all scores of all examined conditions is evaluated at once instead of condition-wise.

3.1. Experimental Set-up

Samples are derived into each condition based on long-duration and clean samples of the I4U file list, prepared for sites participating in NIST SRE'12 [18], by truncation into duration groups of 5 s, 10 s, 20 s, 40 s, and full (original duration) as in [7], 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. The calibration parameters are trained on I4U development set and tested on I4U evaluation set. In total 55 conditions were examined, cf. Tab. 1. The Voice Activity Detection (VAD) labels from clean condition are then applied to corresponding noise versions. Furthermore, in calibration experiments with QMFs including measured SNR, we employed *applied SNR* in order to alleviate deficiencies in SNR estimation, specially in low-SNR conditions. We assume perfect VAD for this experiments in order to exclude undesirable effects rising from VAD shortcomings in low SNRs. While in many scenarios reference samples can be captured under very good conditions, probe samples are usually affected by signal degradation, hence emphasis is put on condition-variable probe samples.

For the sake of tractability of analysis, we experiment only with male speaker data. The i-vectors are compared by PLDA [10] with 200 speaker factors. PLDA is trained in a multi-condition pooled fashion as in [22]. All q-vectors are derived from the original 400-dimensional i-vector space.

3.2. Evaluation criteria

As an application-independent performance metric, we use minimum cost of log-likelihood ratio (LLR) scores $\text{C}_{\text{llr}}^{\text{min}}$, 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 $\mathrm{C_{llr}}$ [24] is computed over genuine and impostor scores S_G, S_I by:

$$\mathrm{C_{llr}} = \sum_{g \in S_G} \frac{\mathrm{ld}(1+e^{-g})}{2\,|S_G|} + \sum_{i \in S_I} \frac{\mathrm{ld}(1+e^{i})}{2\,|S_I|}. \quad (10)$$

The difference of $\mathrm{C_{llr}}$ and $\mathrm{C_{llr}^{min}}$ is referred to as the miscalibration cost $\mathrm{C_{mc}}$ [7]:

$$\mathrm{C_{mc}} = \mathrm{C_{llr}} - \mathrm{C_{llr}^{min}}. \quad (11)$$

Systems of low $\mathrm{C_{llr}^{min}}$ costs are discriminative w.r.t. Bayesian information theory, whereas systems of low $\mathrm{C_{mc}}$ costs are well-calibrated, i.e. $\mathrm{C_{llr}^{min}}$ is well approximated and systems may not need additional information in order to push $\mathrm{C_{llr}}$ further towards $\mathrm{C_{llr}^{min}}$.

3.3. Experimental Results

Baseline results of an uncalibrated system from [8] are shown in Fig. 2: performance in terms of $\mathrm{C_{llr}^{min}}$ degrades significantly in lower SNRs and on shorter observations. All conditions yield respectively high $\mathrm{C_{mc}}$. The gap between $\mathrm{C_{llr}}$ and $\mathrm{C_{llr}^{min}}$ for SNR $\geq 15\,\mathrm{dB}$ is very small.

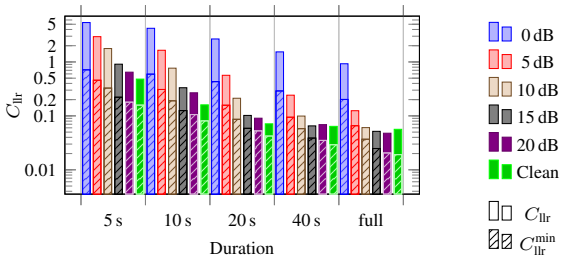
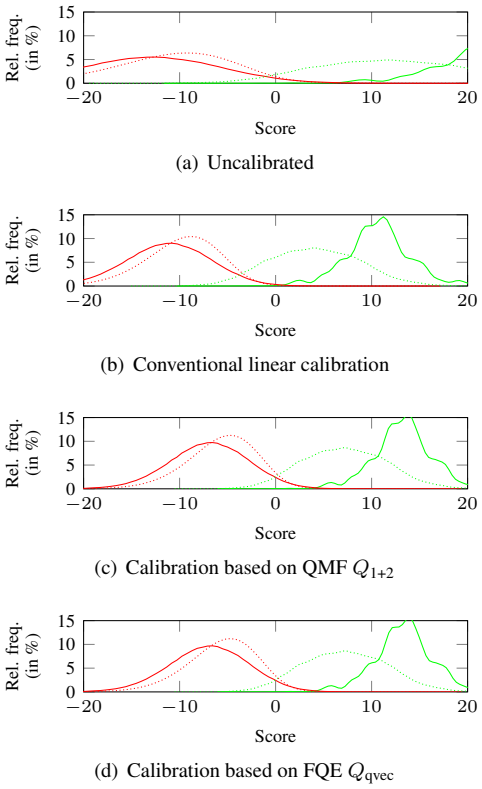


Figure 2: *Baseline performance of uncalibrated system on CROWD conditions, cf. [8].*

Fig. 3 shows effects of conventional, QMF and FQE calibrations regarding the score distribution of genuine and impostor scores. Score distribution alter FQE calibration resembles well the score distribution after QMF calibration. The conventional calibration appears to handle the full/clean condition very well while suffering in variable duration and noise conditions. Alternatively, QMF and FQE show opposite performance: dealing well with noisy and short probe samples. This behavior could be well due to the amount of training data from each condition presented in training process of calibration parameters. Training of conventional calibration is performed using only scores from full/clean condition while for training QMF and FQE scores from 55 conditions are utilized.

3.3.1. *Duration-only Analysis*

In this analysis we focus on noise free probe samples truncated in shorter durations. The calibration parameters for QMF and FQE are trained with recognition scores from noise free probe samples. QMFs and FQE approaches perform better than mismatched calibration on durations$\leq 40\,s$ in terms of $\mathrm{C_{mc}}$, cf. Fig. 4. In terms of $\mathrm{C_{llr}^{min}}$, no significant differences were observed between all examined calibration schemes. When duration of probe samples is $\geq 40\,s$, the probe samples can be considered "full" and the conventional calibration performs best.



(a) Uncalibrated

(b) Conventional linear calibration

(c) Calibration based on QMF Q_{1+2}

(d) Calibration based on FQE Q_{qvec}

Figure 3: *Comparison of genuine (green) and impostor (red) score distributions before and after calibration. Full lines are representing noise free long probe samples. Dashed lines indicate scores from all conditions represented in Tab. 1 excluding full/clean.*

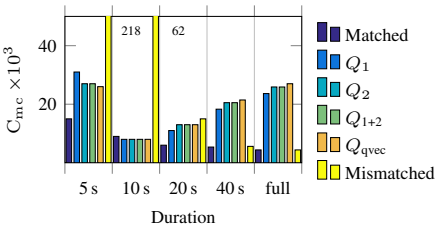
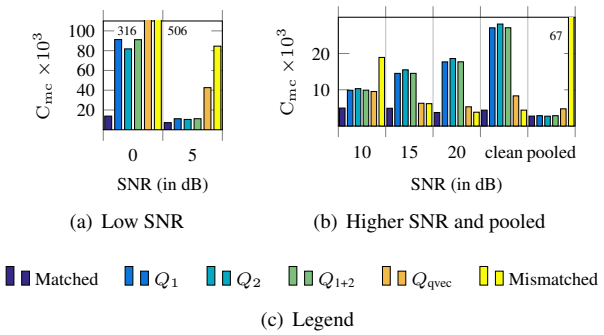


Figure 4: $\mathrm{C_{mc}}$ *comparison on duration variation (clean).*

3.3.2. *SNR-only Analysis*

In this experiment, we consider only noisy probe samples drawn from noise free long duration (full/clean) samples. In training calibration parameters for QMF and FQE, scores from probe samples with no truncation are utilized. In training QMF calibration, applied SNR-levels are used instead of measured SNR. This selection would bias the performance of QMF calibration in different directions depending on SNR region. In low SNR region (SNR $\leq$ 15 dB), estimating SNR-level is problematic due to almost equal level of noise and speech present. Hence, applied SNR-level is much more accurate than a measured SNR-level. On the other side, since original NIST data are seldom noise free, applied SNR-level is less accurate compared to a measured SNR-level for SNR $\geq 10\,\mathrm{dB}$.



(a) Low SNR (b) Higher SNR and pooled

(c) Legend

Figure 5: C_{mc} *comparison on different SNR conditions (full duration).*

Fig. 5 depicts calibration loss, C_{mc}, on variable SNR and full duration condition. In low-SNR region, QMF outperforms FQE based calibration in terms of miscalibration cost, which can be explained by inclusion of applied SNR-level in QMFs. The quality estimates based on UAC aid FQE in handling low-SNR condition compared to the conventional calibration. However, UAC quality estimates are not as accurate as applied SNR-level, which in turn, also undermines FQE calibration training. Dealing with SNR $\geq$ 10 dB, FQE presents superior performance in terms of C_{mc} compared to QMFs. This is in line with the previous argument on SNR estimation in high-SNR, implying that UAC-based quality estimates could be more accurate than applied SNR-level. In terms of C_{llr}^{min}, FQE reveals performance degrades with relative losses to other calibration schemes of 23% – 36% depending on the condition, while other calibration schemes yield almost similar C_{llr}^{min}, cf. Fig. 6.

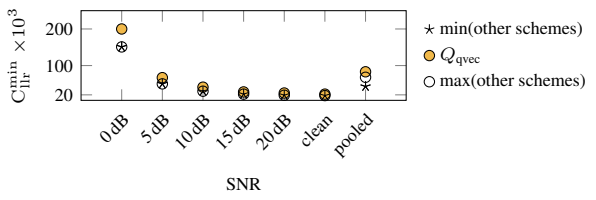


Figure 6: C_{llr}^{min} *comparison of* Q_{qvec} *on SNR conditions to min./max. of other schemes (matched,* $Q_{1,2,1+2}$*, mismatched).*

3.3.3. Combined Duration and SNR effects

Figs. 7, 8 and 9 depict C_{mc} losses for different quality ranges. Thereby, for the sake of tractable analyses, we refer to the term *low quality* if either the duration $\leq$ 10 s or the SNR $\leq$ 5 dB, other conditions are referred to as *good quality*.

In terms of C_{mc}, QMFs and FQE significantly outperform the mismatched calibration scheme on low quality conditions, while by improving signal quality, the mismatched calibration scheme tends to approximate the matched calibration scheme better than QMFs or FQE. For good quality conditions, on increasing signal quality, the calibration loss of QMFs increases, while the mismatched scheme's C_{mc} loss decreases in a continuous fashion. In general, no significant differences are observed between QMFs and FQE in terms of both C_{llr}^{min} and C_{mc}.

From the above analysis we draw the interesting conclusion that UAC-based quality estimates can be successfully applied in calibration stage providing similar level of performance as using oracle SNR-level and duration qualitative in QMF-based calibration.

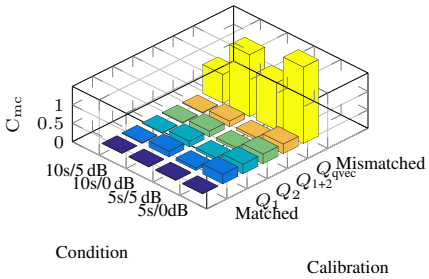


Figure 7: C_{mc} *comparison on combined low qualities.*

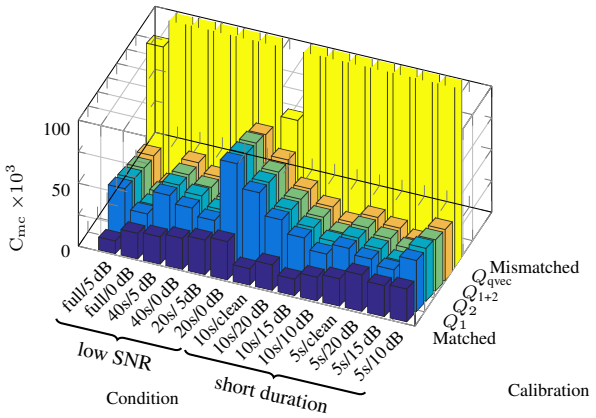


Figure 8: C_{mc} *comparison for having either short probe samples (*$\leq$ 10 *s) or high level of noise (SNR* $\leq$ 5 dB*) present in the probe sample. For the sake of visualizing QMF and FQE behavior,* $C_{mc} > 0.1$ *are cropped.*

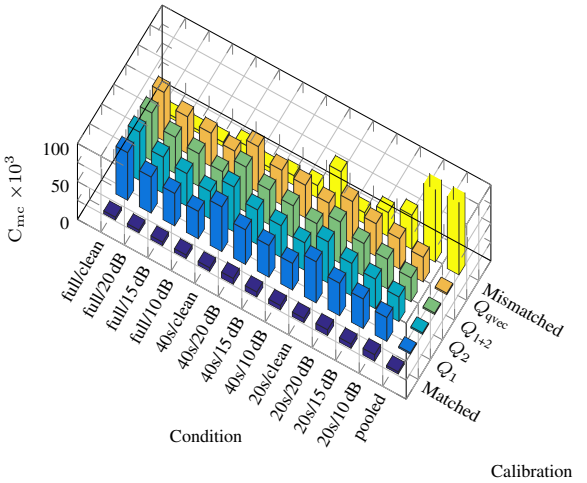


Figure 9: C_{mc} *comparison on good quality probe conditions (duration* $\geq$ 20 *s and SNR* $\geq$ 10 dB*) to pooled condition.* $C_{mc} > 0.1$ *are cropped.*

3.3.4. Analysis of pooled Condition

In Tab. 2, C_{llr}^{min} and C_{mc} performances of conventional, QMFs and FQE calibration approaches are compared for pooled condition with respect to duration-only, SNR-only and combined duration and SNR effects. On pooling variable duration conditions, QMFs and the proposed FQE yield very similar performance in terms of C_{llr}^{min} and C_{mc}. However, by calibrating a pool of scores originated from different SNR levels in probe samples, the presented UAC-based calibration approach in Eq. 9 fails to produce a better performance than QMFs. The C_{llr}^{min} provided by Q_{qvec} is even worse than mismatched calibration. Such behavior could result from poor q-vector modeling of different SNR-levels or in part attributed to the suitability of cosine function in the calibration scheme. When the whole range of duration and SNR variation is introduced in calibration training and evaluation, the calibration performance provided by Q_{qvec} is on a par with QMFs. It is deemed that supplying scores from combined duration and SNR variability to the training stage of calibration results in robust calibration parameter estimation.

Table 2: *QMF comparison on pooled conditions: variable duration (D), variable SNR (S) and combined (C) conditions (in* C_{llr}^{min} $\times 10^3$, C_{mc} $\times 10^3$*).*

Pool	Metric	Matched	Q_1	Q_2	Q_{1+2}	Q_{qvec}	Mismatch
D	C_{llr}^{min}	59	70	68	68	67	70
	C_{mc}	4	4	4	4	4	47
S	C_{llr}^{min}	43	67	68	67	83	68
	C_{mc}	3	3	3	3	5	67
C	C_{llr}^{min}	126	160	160	160	159	160
	C_{mc}	3	2	2	2	2	266

3.3.5. Robustness of Calibration regarding unseen Conditions

Experiments towards robustness are conducted regarding low-quality conditions, i.e. every condition afflicted by 5 s, 10 s or 0 dB, 5 dB is excluded from calibration training. Thus, calibration functions are assumed to be solely trained on biometric data of good quality and sustainable amounts of observations in terms of sample duration. In examining the calibration methods, the whole range of duration and SNR conditions (including low-quality) is tested.

Tab. 3 provides the calibration performance for each experimental set-up on pooled scenarios. In general, QMFs and FQE are still outperforming the mismatched scheme, although C_{mc} costs increased significantly: on duration-only, Q_{qvec} yields the lowest C_{mc}, also yielding the lowest C_{llr}. On noise-only, C_{mc} costs of QMFs are slightly lower than on FQE, gains in terms of C_{llr}^{min} are observed on Q_{qvec}. On combined conditions, C_{mc} costs of the proposed FQE are less affected than on QMFs.

In order to gain more insight towards the robustness issue, more detailed analysis is performed by looking into the individual duration/SNR conditions. Figs. 10, 11 and 12 depict the C_{mc} costs of a calibration training, which is unaware of low-quality conditions. Miscalibration cost, as it was expected, increases on low quality and decreases on higher quality conditions. QMFs and FQE still outperform the mismatched scheme on all low quality conditions. On testing with good quality data, the performance of both QMF and FQE calibration schemes occasionally falls behind the mismatch approach.

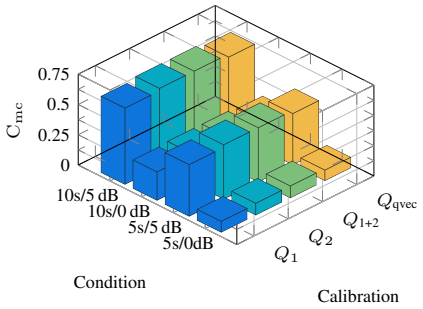


Figure 10: C_{mc} *comparison on combined low-quality, which is excluded from calibration training.*

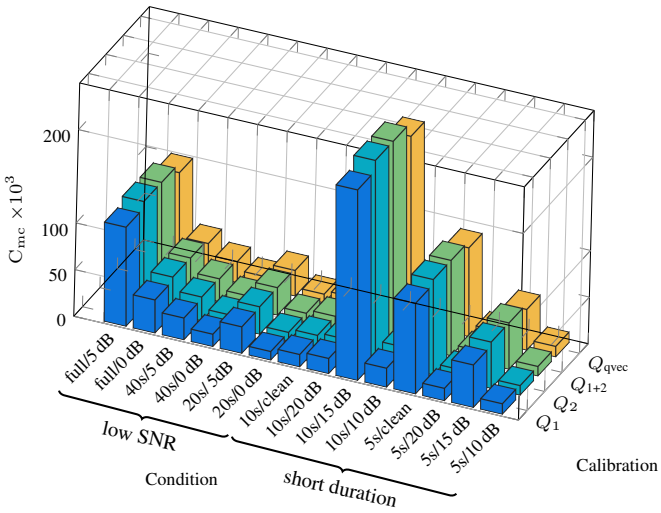


Figure 11: C_{mc} *comparison for having either short probe samples (*≤ 10 *s) or high level of noise (SNR* ≤ 5 dB*) present in the probe sample with excluded low-quality conditions from calibration training.*

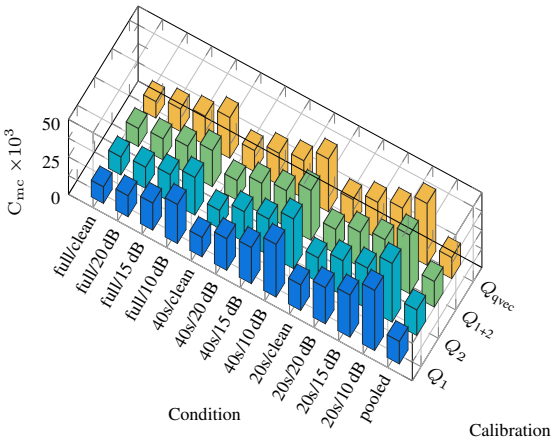


Figure 12: C_{mc} *comparison on good quality probe conditions (duration* ≥ 20 *s and SNR* ≥ 10 dB*) to pooled condition with excluded low-quality conditions from calibration training.*

Table 3: *QMF robustness comparison on pooled conditions with limited data during calibration training: variable duration (D), variable SNR (S) and combined (C) conditions (in* $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ $\times 10^3$, C_{mc} $\times 10^3$*), with differences to Tab. 2 in brackets.*

Pool	Metric	Q_1	Q_2	Q_{1+2}	Q_{qvec}
D	$\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$	69 (-1)	69 (+1)	69 (+1)	67 (0)
	C_{mc}	23 (+19)	23 (+19)	22 (+18)	19 (+15)
S	$\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$	68 (+1)	68 (0)	67 (0)	73 (-10)
	C_{mc}	35 (+32)	36 (+33)	37 (+34)	38 (+33)
C	$\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$	160 (0)	160 (0)	160 (0)	159 (0)
	C_{mc}	15 (+13)	17 (+15)	17 (+15)	14 (+12)

4. Conclusions

Quality-based calibration schemes are essential in facing unconstrained quality conditions, since they are more robust towards unknown conditions than the conventional condition-mismatched calibration approach. A proper function is required to account for single-valued conventional quality measures or vector-based quality estimates in the calibration process. In this paper, we investigated on the performance of q-vector-based calibration using cosine function compared with the performance of recently proposed QMFs for variable utterance duration and noisy conditions. The current study indicates that quality based calibration provides more reliable calibrated scores especially in facing low-quality utterances. More research is in place to find better signal quality modeling and respective handling in the calibration function. Future investigations may also concern impacts of unseen noise types as well as other sample quality and completeness factors e.g., cross-language, cross-channel, reverberant speech and vocal effort (Lombard) effects.

5. Acknowledgment

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).

6. References

[1] S.O. Sadjadi, T. Hasan, and J.H.L. Hansen, "Mean Hilbert Envelope Coefficients (MHEC) for Robust Speaker Recognition," in *Proc. Interspeech*, 2012, pp. 1696–1699.

[2] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2013)*, 2013, pp. 6788–6791.

[3] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 293–298.

[4] R. Saeidi, R. Fernandez Astudillo, and D. Kolossa, "Uncertain LDA: Including observation uncertainties in discriminative transforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press.

[5] S. Cumani, "Fast Scoring of Full Posterior PLDA models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2036–2045, 2015.

[6] R. Saeidi and P. Alku, "Accounting For Uncertainty of i-vectors in Speaker Recognition Using Uncertainty Propagation and Modified Imputation," in *Proc. nterspeech*, 2015, pp. 3546–3550.

[7] 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," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[8] A. Nautsch, R. Saeidi, C. Rathgeb, and C. Busch, "Analysis of mutual duration and noise effects in speaker recognition: benefits of condition-matched cohort selection in score normalization," in *Proc. Interspeech*, 2015, pp. 3006–3010.

[9] M. McLaren, A. Lawson, L. Ferrer, N. Scheffer, and Y. Lei, "Trial-Based Calibration for Speaker Recognition in Unseen Conditions," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.

[10] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] L. Ferrer, L., O. Plchot, and N. Scheffer, "A Unified Approach for Audio Characterization and its Application to Speaker Recognition," in *Proc. Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012, pp. 317–323.

[13] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," in *Conversational Speech, Digital Signal Processing*, 2000, vol. 10, pp. 19–41.

[14] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms," Tech. Rep., Centre de recherche informatique de Montréal (CRIM), 2005.

[15] P.-M. Bousquet, J.-F. Bonastre, and D. Matrouf, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part II*, vol. 8259, chapter Identify the Benefits of the Different Steps in an i-Vector Based Speaker Verification System, pp. 278–285, Springer Berlin Heidelberg, 2013.

[16] S. Cumani and P. Laface, "Generative pairwise models for speaker recognition," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 273–279.

[17] M.I. Mandasari, R. Saeidi, and D.A. van Leeuwen, "Quality measures based calibration with duration and noise dependency for speaker recognition," *Speech Communication*, vol. 72, pp. 126–137, 2015.

[18] R. Saeidi, K. A. Lee, and T. Kinnunen et al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech*, 2013, pp. 1986–1990.

[19] N. Brümmer, D.A. van Leeuwen, and A. Swart, "A comparison of linear and non-linear calibrations for speaker recognition," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 14–18.

[20] N. Bümmer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing," Tech. Rep., AGNITIO Research, South Africa, 2011.

[21] T. Hasan, R. Saeidi, J.H.L. Hansen, and D.A. van Leeuwen, "Duration Mismatch Compensation for i-Vector based Speaker Recognition Systems," in *Int. Conf. on Audio, Speech and Signal Processing (ICASSP 2013)*, 2013, pp. 7663–7667.

[22] 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 *Proc. Int. Conf. on Acoustic, Speech and Signal Processing (ICASSP 2012)*, 2012, pp. 4257–4260.

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

[24] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2–3, pp. 230–275, 2006, Odyssey 2004: The Speaker and Language Recognition Workshop.