Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions

Miranti Indar Mandasari, *Student Member, IEEE*, Rahim Saeidi, *Member, IEEE*, Mitchell McLaren, *Member, IEEE*, David A. van Leeuwen, *Member, IEEE*

***Abstract*—This paper investigates the effect of utterance duration to the calibration of a modern i-vector speaker recognition system with probabilistic linear discriminant analysis (PLDA) modeling. A calibration approach to deal with these effects using *quality measure functions* (QMFs) is proposed to include duration in the calibration transformation. Extensive experiments are performed in order to evaluate the robustness of the proposed calibration approach for unseen conditions in the training of calibration parameters. Using the latest NIST corpora for evaluation, results highlight the importance of considering the quality metrics like duration in calibrating the scores for automatic speaker recognition systems.**

***Index Terms*—calibration, quality measures, duration, forensics, speaker recognition, i-vector, PLDA.**

I. Introduction

The traditional challenges associated with speaker recognition system can be attributed to the *within-speaker variability* of recorded speech signals. Within-speaker or intra-speaker variability [1] refers to the changes that occur in the recorded speech produced by a single speaker. In speaker recognition, the source of within-speaker variability may originate from the language spoken by the speakers, speech register, vocal effort, emotion, background noise, duration of speech samples, recording channel and encoding, and the reverberation conditions. The within-speaker variation has been shown to reduce the performance of speaker recognition system [2]–[7].

In the real application of speaker recognition, there is a high likelihood of having different conditions between the reference (or *model*) and *test* recordings. For example, in a forensic scenario, the test recording might originate from a wire-tapped telephone conversation with the reference speech recorded in the interview session. Another example is in biometric authentication where differences may occur in the reverberation and/or background noise conditions between the enrollment of the speaker and actual authentication attempts.

There are a number of ways in dealing with the problem of within-speaker variability in speaker recognition. Since the Gaussian mixture model (GMM) was proposed for text-independent speaker recognition in the 1990s [8], there has been a strong focus on channel compensation and normalization strategies in feature, score and model domains [9]–[15]. These strategies were proposed to improve system robustness to the within-speaker variability problems.

Along with the development of speaker recognition technology, short duration cases have always been one of many problems that lead to the system performance degradation. As the speech duration is reduced, the system performance tends to follow suit. This is due to the lack of information provided by the short duration of speech samples. In [3]–[5] for example, we can find related studies to discrimination performance of speaker recognition systems in short duration conditions. Even though it is reported in [16] that the i-vector system performance is less sensitive to short utterances compared to previous techniques such as support vector machine (SVM) and joint factor analysis (JFA), performance still degrades in the presence of short duration as presented in [3], [5].

There has been numerous studies in the speaker recognition field in order to solve the short duration problem. In [17], the duration variability problem in speaker recognition is tackled using the duration pattern extracted from the automatic speech recognition prior to the modeling and scoring process. In [18], the short duration problem is addressed by doing logistic regression and fusion from several speaker recognizers.

Almost invariably, the research studying the effects of (shorter) duration in speaker recognition have concentrated on the consequences to the *discrimination* performance, which can be seen from the reported the performance in terms of the calibration-insensitive equal error rate or minimum decision cost function. However, for deployment of speaker recognition systems, the *calibration* of the scores is equally important [20]. Traditional understanding of calibration is the capability of the system to choose a threshold for detection optimally in terms of minimum expected costs. However, in the last decade the concept of calibration has been generalized to a wider range of the detection-error trade-off [21]–[23] with the introduction of the calibrated likelihood ratio and accompanying evaluation metrics such as C_{llr} [22] and the empirical cross entropy [24]. Presentation of recognition results in terms of calibrated log-likelihood-ratios is not only required for application in forensic evidence evaluation [25], but also presents a speaker comparison result in an application-independent way to the user [22], [23]. For the first time in the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluations (SRE) in 2012 [26] it was required to submit a recognition score as a calibrated log-likelihood-ratio.

In our previous work [3], we evaluated i-vector based speaker recognition system with LDA modeling in terms of both discrimination and calibration performances on various duration conditions. However, in that study, we did not propose

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement number 238803.

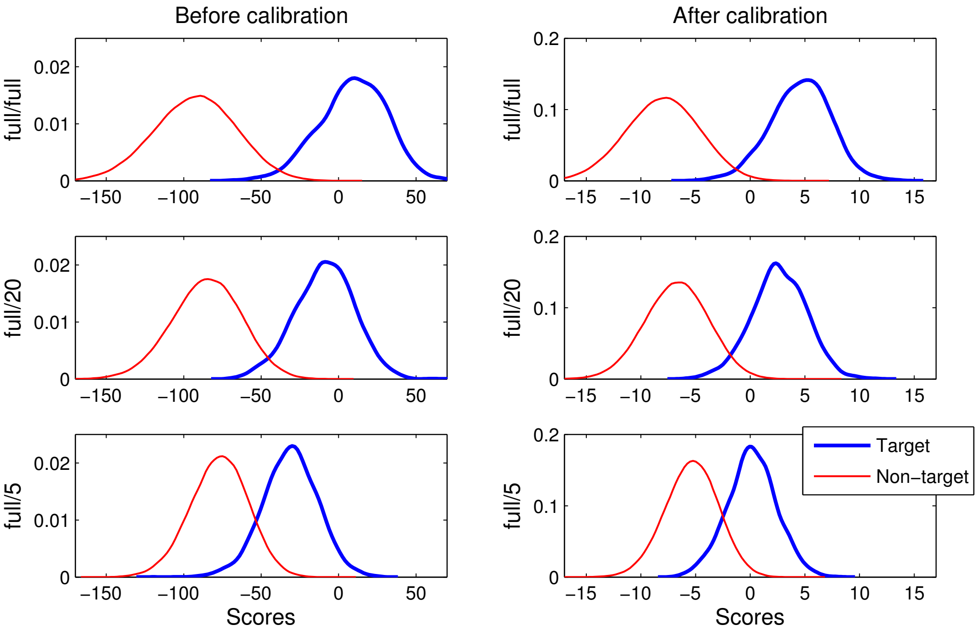


Fig. 1. Score distributions of NIST SRE'10 det-5 database for before and after conventional linear calibration [19] performed in the model/test (sec) duration conditions.

any technique to address this duration variability problem. This paper is a continuation of that work in which we propose a way to address the variability problem during calibration of the speaker recognition system.

The proposed calibration technique that is introduced in this paper is inspired by the concept of *quality measures* presented in [27], [28]. Here, the quality measure is defined as "*knowing the quality of what you have*," which in our case are the speech segments used for speaker recognition. We treated the duration as not only the source of the within-speaker variability, but also as the quality measure or quality factor of speech samples.

Using quality measures of speech to improve the system performance is not a new thing in the speaker recognition field. It is mentioned in [28] that there are four stages in recognition process where the engagement of quality measures is potentially possible in order to improve the system performance: feature extraction, model training, score computation and score fusion. In [29], the quality measures are incorporated in a speaker recognition system in the modeling stage. Here, the authors introduced a technique for combining quality measure information in the system by classifying trials based on speaker adaptation transforms from an automatic speech recognition, and training fusion separately for each of those trial class. The incorporation of quality measures in the score fusion is quite popular in the field with most studies focusing on bi-modal person recognition[1] [30]–[33].

In this paper, we use the duration of model and test segments of speech as the quality measures to improve the *calibration* performance of the speaker recognition system in various duration conditions. As can be seen from Figure 1, the duration variability in speaker recognition system affects the distribution of scores. By keeping the model segment duration as *full* and reducing the duration of the test segment, the target scores distribution approaches the non-target scores distribution (see the before calibration column). By training a typical score calibration technique [19] on *full* duration segments for model and test, we arrive at the right column in Figure 1. When we calibrated the scores from shorter duration conditions using the parameters trained from the longer duration, the large score shift between training and evaluation materials in calibration causes large miscalibration cost. One way of dealing with the score shift in calibration is by using discrete classes for the quality conditions, and effectively training separate calibration parameters for any of the possible combinations of quality conditions between train and test. This was, for instance, carried out for the NIST SRE 2008 by several groups [34], [35] in a calibration implementation coined bi-linear fusion of side-information. These ideas materialized later in the well-known BOSARIS toolkit [36] that can be used for calibrating speaker recognition scores with such side-information. This side-information can be used for quality measures, but inherently as discrete classes.

The new approach taken in this paper is that we model the effect of continuous quality measures to the calibration in low-parameter continuous functions. This is an attempt to capture the relation between a range of quality measure values and the calibration process in a single function, with the potential to both interpolate and extrapolate unseen quality measure values and model the interaction between quality measurements from train and test. We named this proposed calibration technique *Quality Measure Function* or QMF calibration. Please note that we use duration as an example quality measure, but that the approach can also be applied for other

[1] Person recognition based on two biometric modalities (speech, face, fingerprints, etc.).

measures, such as the signal-to-noise ratio. In this paper, we present the results from a number of linear calibration experiments in various duration conditions on a modern *i-vector* based speaker recognition system with *probabilistic linear discriminant analysis* (PLDA) modeling [37]–[39]. Besides the proposed QMF calibration, we also report the calibration performance using other linear calibration techniques such as matched, mismatched, stacked scores, and shared scaling, as comparison to QMF calibration. The proposed approach does not only show improved performance in dealing with duration variation of speech utterances, but also shows some robustness in calibration towards extrapolated durations.

This paper provides an overview of automatic speaker recognition system configuration used for the experiments in Section II. Databases explanation and calibration performance metrics are presented in Section III. In Section IV, all linear calibration approaches analyzed in this paper are explained. The experiment results discussed in Section V, and Section VI concludes the paper.

II. Automatic Speaker Recognition System

Text-independent speaker recognition system technologies have consistently been improving in the past decades [40]. Speaker recognition systems based on Gaussian mixture model (GMM) speaker modeling were proposed in 1995 [8], and became a fundamental approach for speaker recognition with the introduction of the universal background model (UBM) around 2000 [41]. Several milestones in the GMM-UBM based system development were achieved by the researchers subsequently. Support vector machines (SVM) [42] and joint factor analysis (JFA) [43] techniques were introduced from 2003–2007. Both of these are examples of supervector approaches [40]. Recently, the mainstream in the text-independent based speaker recognition system has moved more towards compact representations of the utterance in subspaces, known as i-vectors [16].

The text-independent speaker recognition system used in this paper is based on subspace modeling of i-vectors using probabilistic linear discriminant analysis (PLDA). This section presents a brief explanation of i-vector extraction and PLDA modeling.

A. I-vectors

The speaker recognition system used in this paper follows the i-vector framework that was proposed in [12], [16]. The i-vector is a compact representation of the speech utterance in a low-dimensional space. This space contains both speaker and channel/session variability so that our speaker- and session-dependent Gaussian mean supervector $\mathbf{M}$ can be modeled as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where $\mathbf{m}$ is the speaker- and session-independent mean supervector of the UBM, $\mathbf{T}$ is a low-rank matrix that defines the low-dimensional space, and $\mathbf{w}$ is our identity vector or so-called i-vector.

The speaker- and session-dependent mean supervector in i-vector speech representation is very similar to in the JFA speaker representation [44]. The main difference between the i-vector and JFA modeling is that JFA defines separate speaker and session subspaces, while these factors of variability are combined in a single space $\mathbf{T}$ in i-vector representation.

B. Probabilistic Linear Discriminant Analysis

Probabilistic linear discriminant analysis is a probabilistic approach that models the i-vectors distribution with a Gaussian assumption [37]–[39]. Computed scores from the PLDA model are directly in the form of a ratio of the likelihoods that the enrollment and test i-vectors come from the same speaker and different speakers, respectively. The PLDA method implemented in our system is similar to the approach in [45].

The PLDA models the distribution of i-vectors as the sum of Gaussians for the speaker-dependent term, $\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{y}_k$ and an utterance dependent term $\boldsymbol{\Gamma}\mathbf{z}_r + \boldsymbol{\epsilon}_r$ with $r = 1, \ldots, R$ utterances for a speaker k [15], [37]. The overall mean of the training vectors is denoted by $\boldsymbol{\mu}$ and the matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ are composed of the bases for between-speaker and within-speaker subspaces, respectively. The $\mathbf{y}_k$ and $\mathbf{z}_r$ are positioning the i-vector in between-speaker and within-speaker subspaces, respectively, and $\boldsymbol{\epsilon}_r$ is a Gaussian residual error term with covariance $\boldsymbol{\Sigma}$.

In the context of PLDA model, the hypothesis testing becomes evaluation of the probabilities if the two i-vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, traditionally named enrollment/model and test, are generated by the same speaker, H_1, or by different speakers, H_2. This can be formulated as:

$$s = \frac{P(\mathbf{w}_1, \mathbf{w}_2 | H_1)}{P(\mathbf{w}_1, \mathbf{w}_2 | H_2)} \tag{2}$$

It is shown in [46] and [15] that the likelihoods can be computed analytically as:

$$s = \frac{\mathcal{N}(\mathbf{w}_{12} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_p)}{\mathcal{N}(\mathbf{w}_{12} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_d)}, \tag{3}$$

where $\mathbf{w}_{12}$ is formed by stacking i-vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ and $\boldsymbol{\mu}_2$ by stacking $\boldsymbol{\mu}$ twice, and the covariance matrices for the same and different speakers are obtained by using the matrix expressions:

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Sigma} & \boldsymbol{\Phi}\boldsymbol{\Phi}^T \\ \boldsymbol{\Phi}\boldsymbol{\Phi}^T & \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Sigma} \end{bmatrix} \tag{4}$$

$$\boldsymbol{\Sigma}_d = \begin{bmatrix} \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Sigma} \end{bmatrix} \tag{5}$$

C. Experimental Setup

Features were extracted from audio samples by calculating 19 MFCC[2] parameters and the log energy of speech signal using 20 ms analysis windows every 10 ms. These were augmented using delta and double delta coefficients over 9 frames forming a feature vector of 60 dimension. Speech activity detection (SAD) is performed using a two-Gaussian energy based algorithm as described in [13] and [47]. After SAD, short time Gaussianization is applied using a 5 second

[2] Mel frequency cepstral coefficients.

TABLE I
General System Performance of NIST SRE-2008 and NIST SRE-2010 for male gender in terms of $E_=$ (%).

	$E_=$ (%) for NIST SRE-2008, det-7 (NIST SRE-2010, det-5)				
Model/Test	5	10	20	40	full
5	22.79 (23.33)	18.31 (19.05)	13.27 (14.66)	11.69 (12.33)	10.62 (10.93)
10	16.14 (18.36)	11.07 (13.17)	7.20 (9.36)	6.00 (7.52)	5.47 (6.17)
20	11.28 (14.49)	6.91 (9.48)	4.25 (6.24)	2.97 (4.87)	2.97 (3.81)
40	7.96 (11.66)	5.18 (7.01)	3.06 (4.40)	2.39 (3.35)	1.74 (2.48)
full	7.09 (11.09)	4.27 (5.71)	2.53 (3.48)	1.86 (2.55)	1.33 (1.87)

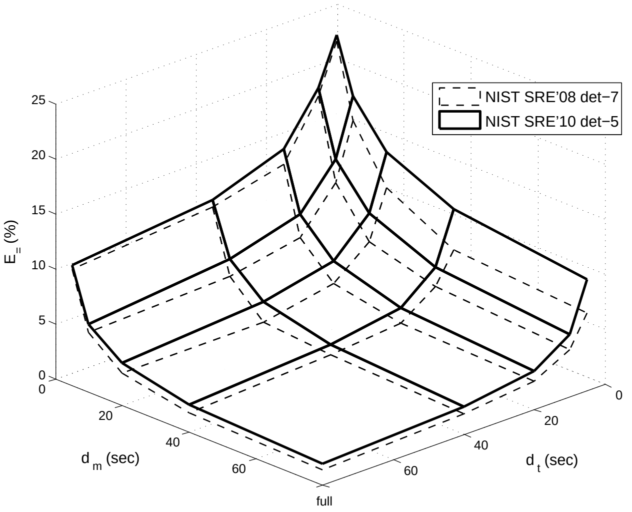


Fig. 2. $E_=$ (%) from NIST SRE-2008 det-7 (dashed lines) and NIST SRE-2010 det-5 (solid lines) databases for male trials in all 25 duration conditions.

analysis window [10]. Finally, a gender-dependent UBM of 2048 components were applied. The UBM was trained on the NIST SRE-2004, 2005, and 2006, Switchboard II and Switchboard Cellular (1 & 2) and Fisher English databases.

In this paper, we used a gender-dependent 400-dimensional i-vector space which was trained on the same data as the UBM training. I-vector length normalization [14] and within class covariance normalization (WCCN) [11] were applied prior to PLDA[3] for optimal performance of our system [2]. We used optimal settings of 200 *speaker factors* and 50 *session factors* in applying PLDA[4]. Subspace matrices in PLDA for both speaker and session spaces are trained using the same databases for i-vector space training, this time using the speaker labels.

III. Evaluation Databases and Metrics

A. NIST Speaker Recognition Evaluation Protocols

For over one decade, the National Institute of Standard and Technology (NIST) have set the standard for evaluation of text independent speaker recognition systems. The general goal of the NIST SRE is to push the technology in the field of text independent speaker recognition forward. At regular intervals, a number of research groups participate with their most advanced technology in speaker recognition, and disclose their findings in the workshop following the evaluation. [48]

In our experiments, we used data and protocols from NIST SRE-2008 [49] and 2010 [50]. We focus on utterances from telephone-telephone conversation in English, which are known as 'det-7' and 'det-5' conditions in SRE-2008 and 2010, respectively. Calibration performance is evaluated on the SRE-2010 trials (extended list) with the calibration parameters trained on the SRE-2008 trials. The experimental results presented in this paper concentrate on male trials only. The number of trials we used from SRE-2008 are 769 target and 10 050 non-target trials, and for SRE-2010, 3 601 target and 226 818 non-target trials, respectively.

B. Utterances Duration and Truncation Procedure

In the NIST SRE database, the length of utterances vary in duration. In order to obtain segments for short duration conditions, all utterances from the database were truncated to $d = 5, 10, 20,$ and 40 seconds. The truncation process was carried out from the beginning point of the utterances at the feature level after SAD and before short term Gaussianization, so that the duration d represents the length of active speech from the utterances. Utterances that have active speech duration less than 40 seconds were excluded from the experiments in order to have the same number of trials in every duration condition.

The original segments from the NIST SRE database without any truncation form the *full* condition in this paper. From the full condition features and the features obtained by truncation, we have five test sets with different duration conditions in both model and test segment collections. Twenty five trial lists are formed by combining the model and test sides from every duration condition for both SRE-2008 and 2010 data sets. This set of 25 trial lists is often referred to as *25 duration conditions* in this paper. We use the notation '$\langle$*duration of model segment*$\rangle$/$\langle$*duration of test segment*$\rangle$ condition,' in which duration is measured in seconds or 'full'.

C. General Discrimination Performance

We have measured the discrimination performance of our PLDA based i-vector system on both the SRE-2008 and 2010 core condition, for trial sets of telephone channel, male speakers. The system's discrimination performance in terms of equal error rate[5] $E_=$ is presented in Table I and depicted in Figure 2. In general, the system shows lower $E_=$ for SRE-2008 than for SRE-2010, in all duration conditions.

[3] We did not apply LDA prior to PLDA modeling.

[4] The dimension of speaker and session factors are the number of components in $\mathbf{\Phi}$ and $\mathbf{\Gamma}$, respectively. See Section II-B for further explanations.

[5] Equal error rate is the error rate at the operating point of a detection system where the probability of false acceptance and probability of false rejection are equal.

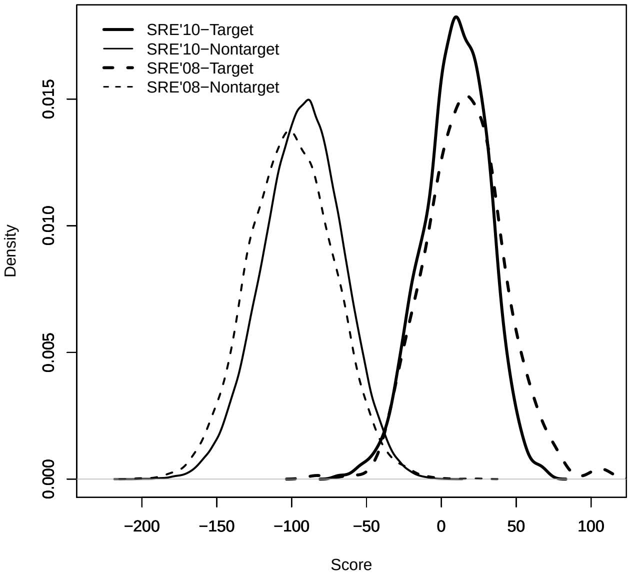


Fig. 3. Score distribution for NIST SRE-2008 ($\mu_{\text{tar}} = 5.8 \pm 24.5$, $\mu_{\text{non}} = -95.2 \pm 26.1$) and NIST SRE-2010 ($\mu_{\text{tar}} = 0.6 \pm 20.7$, $\mu_{\text{non}} = -88.8 \pm 24.6$) of target ($s_{\text{tar}}$) and non-target ($s_{\text{non}}$) uncalibrated scores.

In our previous work [3], we have shown that our i-vector based speaker recognition system has a symmetrical behavior regarding the duration of trials, which is what is expected because model and test segments in the i-vector framework are treated completely the same and the scoring is symmetrical. However, in the most extreme difference between model and test segment durations (5/full and full/5 conditions), we observe a little difference in $E_{=}$ between those conditions. We surmise that this phenomena occurs because of the way NIST decides which utterances are the part of the model or the test segments in their evaluation protocols.

Figure 3 presents the target and non-target scores distribution for SRE-2008 and 2010 databases for full/full duration condition. The figure shows that the scores distribution between the two databases are fairly similar. However, the scores shift from one database to the next, this is a phenomenon known as 'data set shift' [51]. This score shifting will result in lower calibration performance, when using one set (SRE-2008) for calibration of the other (SRE-2010). We may therefore expect some calibration loss using any form of calibration of scores, in this paper we restrict ourselves to linear calibration of scores.

D. Evaluation Metrics for Calibration Performance

To evaluate the calibration performance of the speaker recognition system in general, we use two basic measures C_{llr} and $C_{\text{llr}}^{\text{min}}$. The metric C_{llr} is the cost of the log-likelihood-ratio, a metric that measures calibration over the entire range of effective priors, which has both an interpretation in terms of detection cost functions $C_{\text{det}}(\mathcal{O}_{\text{eff}})$, where $\mathcal{O}_{\text{eff}}$ are the effective prior odds, and an information theoretical interpretation [22]. The metric $C_{\text{llr}}^{\text{min}}$ is the same metric, but after an optimal transformation of scores that minimizes C_{llr} under the condition that the order of scores stay the same, i.e., the score to likelihood function is monotonously rising. From these basic metrics, we derive the absolute and relative *miscalibration costs*, or calibration loss. For an introduction to C_{llr}, see [23].

The metric C_{llr} can be evaluated empirically for a supervised set of evaluation log-likelihood-ratios x_i (in our case SRE-2010) using

$$\begin{aligned} C_{\text{llr}} =& \frac{1}{N_{\text{tar}}} \sum_{i \in \text{tar}} \log_2(1 + \exp(-x_i)) \\ &+ \frac{1}{N_{\text{non}}} \sum_{j \in \text{non}} \log_2(1 + \exp(x_j)) \end{aligned} \tag{6}$$

with x_i and x_j running over the number of target trials (N_{tar}) and non-target trials (N_{non}) respectively, i.e., trials for which either H_1 or H_2 is true.

The absolute (C_{mc}) and relative (R_{mc}) calibration loss, or miscalibration cost, are defined as:

$$C_{\text{mc}} = C_{\text{llr}} - C_{\text{llr}}^{\text{min}}, \tag{7}$$

and

$$R_{\text{mc}} = \frac{C_{\text{mc}}}{C_{\text{llr}}^{\text{min}}} = \frac{C_{\text{llr}}}{C_{\text{llr}}^{\text{min}}} - 1 \tag{8}$$

The value for minimum cost of the log-likelihood-ratios, $C_{\text{llr}}^{\text{min}}$, can be obtained by isotonic regression. An efficient method for this is known as the pool adjacent violators (PAV) algorithm as explained in [22] which has relations to the receiver operating characteristic convex hull (ROC-CH) [52]. All metrics share the same property, that lower values are indicate better performance. Here, C_{llr} integrates both discrimination and calibration performance, where $C_{\text{llr}}^{\text{min}}$ only reveals discrimination performance. The mis-calibration costs C_{mc} and R_{mc} only show calibration performance.

IV. Score Calibration

In many mathematical formulations of speaker recognition, including PLDA based systems, recognition scores are computed as likelihood ratios. However, due to a number of modeling assumptions that are probably incorrect, most notably the assumption of frame independence, these computed scores do not have a direct proper probabilistic interpretation. Using such uncalibrated scores in court as calibrated likelihood ratios will be misleading [53]. However, there is a number of ways in which we can transform the uncalibrated scores into log-likelihood-ratios, a process known as *calibration* and in which the field of speaker recognition has extensive experience, specifically in comparison to other biometric technologies. A remarkable property of calibrated log-likelihood-ratios ℓ is relates to the probability density function of itself (for a derivation of this, see the appendix),

$$\ell = \log \frac{P(\ell \mid H_1)}{P(\ell \mid H_2)}. \tag{9}$$

The metric C_{llr} can measure the validity of this property empirically for a set of evaluation trials.

In this paper we restrict ourselves to *linear* calibration transformations[6], i.e., the function that is used to convert our uncalibrated scores s into calibrated likelihood ratios x is

$$x = w_0 + w_1 s \tag{10}$$

where w_0 is the offset of the transformation and w_1 is a scaling parameter. Both the offset and scaling parameters can be obtained by optimization on a development set. An effective method for this optimization is logistic regression [56] training which uses an objective function quite closely related to C_{llr} [22]. For implementation of this linear calibration, we utilized FoCal toolkit [19] and the `sretools` analysis package [57].

There are 5 linear calibration approaches discussed in this paper which we refer to as 'mismatched,' 'matched,' 'stacked,' 'shared scaling' and 'duration quality measure function' (QMF). All of these approaches are explained in the following subsections. The goal of this research is to describe the problems of duration for calibration (mismatched, matched and stacked scores), to understand the effects of duration to the calibration parameters (shared scaling), and to design low-parameter models to account for these effects in calibration (*duration quality measures*).

A. Calibration using Mismatched and Matched Duration Conditions

Both mismatched and matched approaches employ the scores transformation defined in equation (10), which consists of two weighting parameters w_0 and w_1. In the mismatch approach, the two calibration parameters are trained in the full/full duration condition from the SRE-2008 calibration set. These parameters are then applied to all 25 duration conditions of the SRE-2010 evaluation set. This approach is called *mismatched* because the presence of many duration-mismatched conditions between the calibration and evaluation data, i.e., the calibration parameters remain trained on the full/full condition even if they are applied to, e.g., the 40/20 seconds condition for evaluation. This is a 2-parameter calibration, and we expect it to be the worst performing based on previous work on an LDA i-vector system [3].

In the *matched* approach, calibration parameters are trained on each of 25 duration conditions in the SRE-2008 calibration set. This approach uses 50 calibration parameters, 25 pairs of weighting parameters w_0 and w_1. Each of these weighting pairs are then applied to the corresponding matched condition in the SRE-2010 evaluation data. One may consider this as a "poor-man's" solution, because it does not rely on understanding the effect of the quality measures on calibration, but requires to match the quality of the calibration data with the conditions under evaluation. For (shorter) *duration*, this may actually be feasible (although perhaps not very practical), but for other quality factors such as language, reverberation or background noise level and type, this may be less than trivial. In the more realistic forensic cases, it may not always be possible to find training calibration material that match the condition of evaluation data. The *matched* approach needs 50 parameters, and we expect this to outperform most other methods, simply because the only reason for miscalibration is the data set shift, which is hard to deal with anyway.

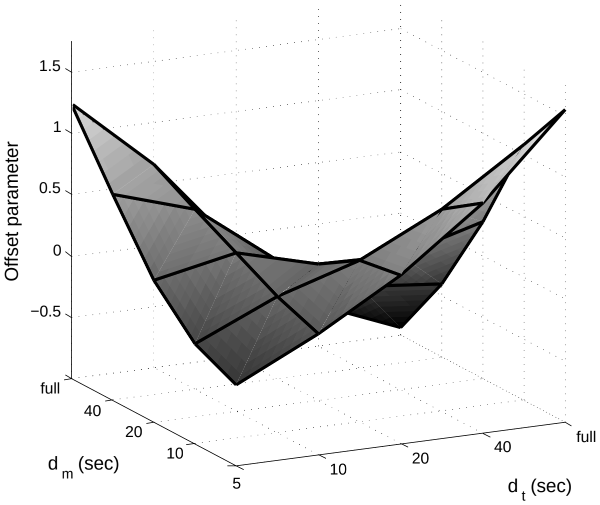


Fig. 4. Offset parameters distribution across all duration conditions from shared scaling calibration approach. This distribution has a saddle-plane shape.

B. Calibration using Stacked Scores

In the case of mismatched duration, the training calibration condition used (full/full) is quite peripheral to the conditions used in evaluation. The purpose of the *stacked scores* approach is to understand the potential of two-parameter linear calibration if calibration training data shows the same variability as the evaluation. The stacked scores approach uses all available data for calibration in all 25 conditions in order to train a single set of transformation parameters according to (10). Similarly to the mismatched approach, this single set of 2 calibration parameters is applied to all duration conditions of the evaluation data. We expect that this approach performs better, on average, than the mismatched approach.

C. Calibration using Shared Scaling

Keeping in mind that we would like to design a calibration function that takes duration of model and test segments into account, we want to study the effect of duration on the shift parameter w_0 while keeping the scaling w_1 constant. This approach is somewhat between stacked scores and matched approaches. We stack all calibration trials, but use a single scaling parameters (w_1) while allowing for a duration-dependent offset parameter (w_0) for each of the 25 duration conditions. This approach is similar to the 'side information' calibration technique mentioned earlier that was employed in the NIST SRE-2008 in order to deal with varying language and

[6] It is presented in [54] that the scores calibration can be conducted by doing simple normalization (like Z- or T-norm), isotonic regression, etc. besides the linear logistic regression. One of the authors is also contributed in the proposing of line-up calibration method for speaker recognition system which is not a form of linear calibration [55].

TABLE II
DURATION QUALITY MEASURE FUNCTIONS (QMFS) PROPOSED FOR CALIBRATION ON VARIOUS DURATION CONDITIONS.

n	QMF: $Q_n(d_m, d_t, \ldots)$	Additional parameters
1	$Q_1 = w_2 \left\vert \log \frac{d_m}{d_t} \right\vert$	w_2
2	$Q_2 = w_2 \log^2 \frac{d_m}{d_t}$	w_2
3	$Q_3 = w_2 \log \frac{d_m}{d_c} \log \frac{d_t}{d_c}$	w_2, d_c
4	$Q_4 = w_2 \left(\log \frac{d_m}{d_c} + \log \frac{d_t}{d_c}\right)^2$ $-w_3 \left(\log \frac{d_m}{d_c} - \log \frac{d_t}{d_c}\right)^2$	w_2, w_3, d_c

TABLE III
MINIMUM COST OF LOG-LIKELIHOOD-RATIO CALIBRATION $C_{\text{llr}}^{\text{min}}$ OF NIST SRE-2010 DET-5 CONDITION FOR THE MATCHED, MISMATCHED, AND STACKED SCORES CALIBRATION TECHNIQUES.

d_m/d_t	5	10	20	40	full
5	0.695	0.581	0.476	0.416	0.372
10	0.572	0.428	0.321	0.264	0.223
20	0.470	0.319	0.219	0.171	0.138
40	0.401	0.248	0.156	0.119	0.094
full	0.351	0.210	0.122	0.090	0.071

transducer [34]. We expect this approach to be almost as good as the matched approach, because of its many parameters. However, this is not an approach we would like to propose as a viable method of dealing with continuous quality factors in general, but rather as an inspiration for designing quality measure functions. To this effect, the 25 offset parameters are presented in Figure 4. The behavior study of the calibration parameters presented in the figure becomes the foundation of the duration quality measures approach that we propose in the next section.

D. Linear Calibration with Duration Quality Measure Functions

Finally, we propose a calibration approach for calibration in various duration conditions that models the effect of calibration in a low-parameter model. The general score transformation model is:

$$x = w_0 + w_1 s + Q(d_m, d_t, w_2, \ldots) \tag{11}$$

where $Q(d_m, d_t, w, \ldots)$ is the quality measure function (QMF) that is related to duration of model segment d_m and duration of test segment d_t. We propose four QMFs for improving the global calibration performance in various duration conditions in Table II.

The results on calibration using shared scaling (cf. Figure 4 (further explained in Section V) shows that the larger magnitude of offset parameters (w_0) occurs where the difference between model and test segments duration are larger. The first two QMFs, Q_1 and Q_2, model this behavior. These two functions, however, do not model any difference in offset where model and test segments have the same duration. Observing Figure 4 there clearly is a dependency on the duration even for $d_m = d_t$. Therefore, the QMFs Q_3 and Q_4 were proposed in order to better model the offset parameters behavior from the shared scaling approach.

The mathematical form of Q_3 and Q_4 are modeled after the saddle-like shape of the surface in Figure 4. In general, a two-dimensional saddle function can be described by

$$f(y, z) = \alpha \cdot y^2 - \beta \cdot z^2, \tag{12}$$

with $\alpha\beta > 0$. In our case, we use a rotated version of the axes y and z which work with log-duration, and placing the origin at $d_c = 20\,\text{s}$ for both model and test durations, so that

$$y = \log \frac{d_m}{d_c} + \log \frac{d_t}{d_c}, \tag{13}$$

$$z = \log \frac{d_m}{d_c} - \log \frac{d_t}{d_c}. \tag{14}$$

The rotation is because the distribution of offset parameters has the saddle plane shape which lies along the diagonal axis of d_m and d_t as depicted in Figure 4. By using y and z defined in (13), the QMFs that model the saddle-plane Q_3 and Q_4 from Table II can be found with

$$w_2 = 2(\alpha + \beta), \tag{15}$$

$$w_3 = \alpha - \beta. \tag{16}$$

The third QMF Q_3 is the case where $\alpha = \beta$, i.e., forcing the 'tails' at the extremes full/5 and 5/full to go as much up as the ones at 5/5 and full/full go down. The parameter d_c is fixed in our experiments to 20 seconds as it is the center of our saddle-shaped parameter distribution (as seen in Figure 3)

The proposed QMFs are not designed to handled the condition where duration goes to zero. In this extreme condition, the speaker recognition system should output $\ell = 0$ as there is no speaker information, and hence both hypotheses are equally likely. The log-duration dependence of the QMFs in Table II is inspired by Figure 4, which has logarithmic axes, and may find some further motivation in the observation that the number of unique phones found in a random speech sample scales logarithmically with duration over a fairly wide range [58].

V. EXPERIMENT RESULTS

The calibration results for all linear calibrations mentioned in previous section are presented in Table IV, and analyzed in the paragraphs V-B until V-E. Table V is the summary of Table IV in where we take the average μ and standard deviation σ across all 25 duration conditions for each calibration technique[7].

A. Discrimination performance based on $C_{\text{llr}}^{\text{min}}$

The minimum achievable values $C_{\text{llr}}^{\text{min}}$ were measured and showed in Table III for all 25 duration conditions in the SRE-2010 evaluation set. The numbers presented in Table III are the $C_{\text{llr}}^{\text{min}}$ values from all calibration techniques but the QMFs approach. The $C_{\text{llr}}^{\text{min}}$ values on the full/full QMFs calibration

[7] We took the averaging approach to summarize the results instead of pooling the scores then computing the calibration metrics. This is due to the pooling method that causing C_{mc} less sensitive because of the pooled $C_{\text{llr}}^{\text{min}}$ is increasing.

TABLE IV
CALIBRATION RESULTS OF LINEAR CALIBRATION APPROACHES IN TERMS OF C_{llr} C_{mc} AND R_{mc} FOR ALL 25 DURATION CONDITIONS.

		C_{llr}					C_{mc}					$R_{\mathrm{mc}}(\%)$				
Calibration:	d_m/d_t	5	10	20	40	full	5	10	20	40	full	5	10	20	40	full
Mismatched	5	.771	.684	.603	.591	.600	.076	.103	.127	.174	.228	10.96	17.73	26.68	41.84	61.19
	10	.668	.521	.398	.346	.308	.096	.093	.078	.081	.086	16.81	21.63	24.26	30.72	38.64
	20	.600	.403	.267	.205	.163	.130	.084	.049	.034	.025	27.59	26.32	22.33	19.77	18.02
	40	.571	.328	.187	.134	.101	.170	.081	.031	.015	.007	42.34	32.49	20.15	12.59	7.67
	full	.568	.294	.145	.098	.074	.217	.084	.024	.008	.004	61.71	40.25	19.32	8.49	5.53
Matched	5	.701	.586	.482	.423	.380	.006	.005	.006	.007	.008	0.82	0.86	1.18	1.64	2.18
	10	.581	.435	.326	.270	.231	.009	.007	.005	.006	.009	1.63	1.59	1.63	2.33	3.94
	20	.480	.324	.223	.176	.144	.010	.005	.005	.004	.005	2.04	1.72	2.23	2.57	3.84
	40	.410	.252	.160	.124	.099	.009	.004	.005	.005	.005	2.25	1.68	2.92	4.27	5.21
	full	.360	.214	.127	.095	.074	.008	.004	.005	.005	.004	2.42	2.01	3.88	5.02	5.53
Stacked scores	5	.752	.595	.484	.437	.415	.057	.014	.008	.021	.043	8.16	2.41	1.67	4.97	11.57
	10	.586	.434	.326	.272	.233	.014	.006	.006	.008	.011	2.46	1.37	1.76	2.92	4.94
	20	.478	.324	.227	.180	.147	.008	.006	.009	.009	.009	1.68	1.75	4.05	5.26	6.15
	40	.421	.256	.168	.134	.112	.020	.008	.013	.016	.018	4.88	3.22	8.14	13.20	19.15
	full	.392	.219	.133	.107	.096	.040	.009	.011	.017	.025	11.42	4.52	9.38	19.20	35.70
Duration QMF Q_1	5	.731	.597	.484	.424	.378	.036	.015	.007	.007	.006	5.21	2.64	1.55	1.78	1.63
	10	.586	.451	.328	.271	.232	.014	.023	.008	.007	.009	2.49	5.41	2.47	2.46	3.85
	20	.477	.328	.232	.178	.151	.007	.009	.013	.007	.011	1.49	2.79	6.16	4.00	8.24
	40	.408	.254	.165	.126	.109	.007	.007	.009	.008	.015	1.71	2.63	5.93	6.48	15.96
	full	.355	.215	.135	.102	.085	.006	.007	.014	.013	.014	1.64	3.34	11.83	14.45	20.63
Duration QMF Q_2	5	.735	.597	.486	.426	.381	.040	.016	.010	.009	.007	5.76	2.71	2.03	2.23	1.82
	10	.586	.440	.329	.272	.232	.015	.012	.009	.008	.007	2.54	2.70	2.68	3.00	3.26
	20	.480	.328	.227	.178	.148	.010	.010	.009	.007	.009	2.06	3.04	3.90	4.02	6.22
	40	.410	.256	.165	.128	.107	.009	.008	.009	.009	.013	2.13	3.16	5.88	7.75	13.85
	full	.356	.214	.132	.102	.087	.006	.006	.011	.012	.016	1.74	3.03	9.51	13.17	23.44
Duration QMF Q_3	5	.740	.603	.485	.424	.379	.045	.022	.009	.008	.007	6.46	3.75	1.89	1.83	1.81
	10	.591	.436	.326	.271	.231	.020	.008	.005	.006	.008	3.44	1.95	1.58	2.42	3.59
	20	.479	.324	.226	.179	.146	.009	.005	.008	.008	.008	1.90	1.55	3.54	4.72	5.59
	40	.408	.254	.167	.129	.102	.007	.006	.011	.010	.007	1.74	2.59	7.36	8.84	7.83
	full	.356	.215	.132	.099	.077	.006	.006	.010	.008	.005	1.67	3.04	8.51	8.32	6.58
Duration QMF Q_4	5	.743	.604	.485	.424	.378	.048	.022	.009	.008	.006	6.94	3.85	1.87	1.84	1.66
	10	.592	.436	.326	.271	.231	.020	.008	.005	.007	.008	3.54	1.79	1.59	2.49	3.45
	20	.479	.323	.227	.180	.146	.009	.005	.008	.009	.007	1.87	1.49	3.88	5.10	5.39
	40	.408	.254	.168	.130	.101	.007	.007	.012	.011	.007	1.75	2.68	7.92	9.28	7.31
	full	.355	.215	.132	.099	.077	.006	.006	.010	.007	.004	1.68	2.94	8.43	8.07	5.86

are slightly different to what is presented in the table due to the effect of variable duration in the full/full condition. As explained in Section III-D, $C_{\mathrm{llr}}^{\mathrm{min}}$ is a representation of discrimination loss [23]. Thus, it has similar information as $E_{=}$ that it is presented in Section III-C. We can observe $C_{\mathrm{llr}}^{\mathrm{min}}$ is increasing as the duration of model/test segments are decreasing. Note that $C_{\mathrm{llr}}^{\mathrm{min}}$ stays the same after linear calibration applied unless $C_{\mathrm{llr}}^{\mathrm{min}}$ is computed on pooled scores over all 25 duration conditions, because the pooling makes $C_{\mathrm{llr}}^{\mathrm{min}}$ sensitive to "relative calibration" between the duration conditions. Further, linear calibration will have an effective objective to minimize C_{llr}, but we know beforehand that $C_{\mathrm{llr}}^{\mathrm{min}}$ is a lower bound to this.

TABLE V
RESUME OF TABLE IV: CALIBRATION PERFORMANCE OVER ALL LINEAR CALIBRATION APPROACHES IN TERMS OF THE MEAN μ AND STANDARD DEVIATION σ OF ALL 25 DURATION CONDITIONS.

Approach	C_{llr}		C_{mc}		$R_{\mathrm{mc}}(\%)$		$n_p^{*)}$
	μ	σ	μ	σ	μ	σ	
Mismatched	.385	.216	.084	.063	26.20	14.86	2
Matched	.307	.173	.006	.002	2.62	1.35	50
Stacked Scores	.317	.177	.016	.013	7.60	7.76	2
Shared Scaling	.313	.173	.012	.005	5.46	3.42	26
Duration Q_1	.312	.175	.011	.007	5.47	5.08	3
Duration Q_2	.312	.175	.011	.007	5.26	5.04	3
Duration Q_3	.311	.177	.010	.008	4.10	2.54	3
Duration Q_4	.311	.177	.010	.009	4.11	2.56	4

$^{*)}n_p$ = number of calibration parameters.

B. Calibration using Mismatched and Matched approaches

In this section, we compare the calibration performance from the mismatched and matched approaches. The calibration results for these two approaches are presented in the top sections of Table IV. In the mismatched condition, where we trained calibration parameters on the full/full duration condition only, the miscalibration values are higher, as the durations of model or test segments are shorter, and deviate more from the calibration condition. The highest miscalibration values are present when there is a large difference between the duration in model and test segments. These results confirm that by training the calibration parameters using a data set which is not representative of evaluation data in the sense of speech segments duration, the calibration performance is not very good.

The matched calibration approach is an easy solution to come up with the deficiency of mismatched approach by simply matching the duration condition of training calibration data to the evaluation data. As presented in Table V, the calibration system performance using matched approach totally surpass the mismatched approach based on the miscalibration values. The average miscalibration values from all 25 duration conditions drops from 0.084 for mismatched to 0.006 for matched. This is of course an expected effect, but it is interesting to note here that the effect on miscalibration cost in the PLDA system is much smaller than in the i-vector LDA system reported on our earlier work [3]. For the LDA system, we had observed average miscalibration values $C_{\rm mc} = 0.494$ and $R_{\rm mc} = 141\,\%$ over all 25 duration conditions following the mismatched calibration approach. This shows that with the classifier in i-vector systems becoming better at discrimination (going from LDA with cosine distance scoring via normalized cosine to PLDA with probabilistic scoring), the calibration behavior improves as well.

Even though the matched calibration approach offers a very good calibration performance, the training process within this approach can be tough because we have to match the condition of our training data to the evaluation data. This condition matching might not possible in an extreme condition such as when we have only limited amount of training data in which we do not have long enough duration to match the evaluation data condition.

C. Calibration using Stacked Scores

The experiment results of the stacked scores approach is presented in the third section of Table IV. Compared to the mismatched approach, the stacked calibration results is generally better in terms of miscalibration results over all 25 duration conditions. The miscalibration metrics in the condition where the duration of model and test segments have large difference (i.e., 5/full and full/5 conditions) are lower in the staked scores approach than in the mismatched approach. These results show that the stacked scores calibration works generally better than the mismatched approach even though we have to sacrifice the calibration performance in the longer duration condition.

The average values of miscalibration using the stacked scores approach is 0.016, still a notable increase over the optimal value of 0.006 found using the matched approach. This is due to a fact that in the stacking calibration, we only use 2 calibration parameters even though we had more training data available compared to the matched approach, while in the matched approach, we use 50 calibration parameters in total. Hence, there is room for improvement for the system calibration performance in various duration conditions, where we can have a good performance with using only a small number of calibration parameters. In order to find out about the pattern of calibration parameters with respect to duration of model/test segments, we will now present the results for *shared scaling* calibration which are discussed in the next section.

D. Calibration using the shared scaling approach

This section presents the experimental results on calibration using the shared scaling approach. This experiment was performed to demonstrate the relation between the bias term (offset parameter) of linear calibration and the speech segment duration. This formed the inspiration of the proposed duration quality measure function which explained in Section V-E. The 25 offset parameters trained from this calibration approach is presented in Table VI and have been shown before in Figure 4. The standard deviation of this 25 offset parameters is 0.511. This means that there is a large variability of offset parameters due to the variation of duration condition in the calibration.

TABLE VI
OFFSET PARAMETERS IN SHARED SCALING APPROACH FOR ALL 25 DURATION CONDITIONS ($\mu = 0.236$, $\sigma = 0.586$).

d_m/d_t	5	10	20	40	full
5	-0.346	-0.022	0.372	0.877	1.552
10	-0.189	0.114	0.322	0.645	1.085
20	0.153	0.294	0.110	0.094	0.285
40	0.680	0.471	-0.037	-0.340	-0.402
full	1.235	0.660	-0.103	-0.669	-0.934

TABLE VII
P-VALUES FROM ONE-SIDED PAIRED T-TEST IN COMPARING $R_{\rm mc}$ VALUES OVER ALL 25 DURATION CONDITIONS FOR CALIBRATION APPROACH IN SIDE-A AND SIDE-B WITH ALTERNATIVE HYPOTHESIS: SIDE-A GIVES "GREATER" $R_{\rm mc}$ VALUES THAN SIDE-B CALIBRATION APPROACH.

side-B	p-values (p) based on $R_{\rm mc}$ metric			
side-A	Q_1	Q_2	Q_3	Q_4
Mismatched	$4.6{\times}10^{-6}$	$3.9{\times}10^{-6}$	$3.8{\times}10^{-7}$	$3.8{\times}10^{-7}$
Stacked scores	$1.1{\times}10^{-2}$	$2.3{\times}10^{-3}$	$6.7{\times}10^{-3}$	$8.0{\times}10^{-3}$
Q_1	-	$2.1{\times}10^{-1}$	$3.3{\times}10^{-2}$	$4.3{\times}10^{-2}$
Q_2	$7.9{\times}10^{-1}$	-	$6.3{\times}10^{-2}$	$7.5{\times}10^{-2}$
Q_3	$9.7{\times}10^{-1}$	$9.4{\times}10^{-1}$	-	$5.5{\times}10^{-1}$
Q_4	$9.6{\times}10^{-1}$	$9.3{\times}10^{-1}$	$4.5{\times}10^{-1}$	-

From Table V, the average of miscalibration values of the shared scaling[8] approach is 0.010 which is better than the stacked scores calibration results. Despite its low average of miscalibration costs, the shared scaling approach requires a large number of calibration parameters (26) which need to be trained. As indicated earlier, the motivation was to use the offset parameters (cf.Figure 4) to find a simpler calibration technique with few calibration parameters and comparable calibration performance. In the next subsection, we present the duration quality measures calibration results.

E. Calibration using the Duration Quality Measures Approach

This section presents the results of duration quality measure calibration which is based on the QMFs proposed in Section IV-D. As can be seen from the average of miscalibration values of all 25 duration conditions in Table V, all of the proposed duration QMFs offer better calibration performance than the stacked scores approach. It has also comparable performance to the shared scaling approach, even though the duration QMF approaches use only 3 or 4 calibration parameters. By adding 1 or 2 extra parameters in the duration QMF approach, it provides much better calibration performance than the 2-parameter mismatched or stacked scores approaches. The differences in $R_{\rm mc}$ are statistically significant at $p < 0.05$, as tested in a one-sided paired t-test[9] [59]. The results for this t-test are presented in Table VII.

Presented in Table V, there are differences in performance which are offered by each of the proposed duration QMFs. The first two functions, Q_1 and Q_2, have similar calibration

[8] The calibration experiment results for *shared scaling* are not shown in Table IV because this calibration method is carried out only as an inspiration for QMF design, and not as a calibration method per sec.

[9] via R programming language.

performance trend in terms of miscalibration values, while the last two functions, Q_3 and Q_4, have almost identical calibration performance trend. In terms of the average of miscalibration values of all 25 duration conditions, Q_3 and Q_4 offer slightly better calibration performance than the Q_1 and Q_2 functions, but it has bigger variance. A one-sided paired t-test shows that Q_3 and Q_4 have statistically significantly lower R_{mc} than Q_1 at the $p < 0.05$ level as presented in Table VII.

In the miscalibration metrics of the proposed QMFs shown in Table IV, Q_1 and Q_2 functions have better calibration performance at the 5/5 duration condition, while it has worse calibration performance at the full/full condition compared to the Q_3 and Q_4 functions. In general, Q_3 and Q_4 functions have lower miscalibration value across all duration conditions except for 3 conditions, which are 5/5, 5/10 and 10/5. From this observation, we summarize that the Q_3 and Q_4 functions have better performance than the Q_1 and Q_2 functions except in the conditions in which the model and/or test segments contain 5 second of duration. Even though there are slight differences between the calibration performance among the proposed duration QMFs, all of them bring improved results in the calibration performance of the system compared to the stacked calibration technique, as revealed by one-sided paired t-tests we carried out.

To better analyzed how the QMFs can give better performance in calibration with duration variability problem, we present the score distributions of calibrated scores from the mismatched and Q_4 calibrations in Figure 5. As can be seen from the figure, the calibrated scores from mismatched technique are shifted further to the left when the test segment duration is decreased. In the calibrated scores using the Q_4 function, however, the QMF is able to normalized the duration effect in the scores distribution. Therefore, the Q_4 calibrated scores is pushed back to the center of the log likelihood ratio (LLR) axis (LLR $= 0$).

We further evaluate the robustness of proposed QMF calibration to the mismatched channel problem. We applied the same calibration parameters used to calibrate NIST SRE'10 det-5 condition to the det-3 condition. Since the calibration parameters were trained in the NIST SRE'08 det-7 (telephone-telephone), the evaluation on NIST SRE'10 det-3 (interview-telephone) is therefore incurring 'mismatched channel' challenge in calibration. The average of C_{llr} values from all 25 duration conditions in det-3 trial set are 0.550 (mismatched); 0.376 (matched); 0.380 (stacked scores); 0.379 (Q_1 and Q_2); and 0.375 (Q_3 and Q_4). The C_{llr} values from det-3 evaluation deviate by similar amount for each of the calibration techniques compared to the det-5 results. In addition, similar trends in performance are observed across all calibration techniques. These results show that the QMF calibration is robust in dealing with duration variability, even though the mismatched channel problem occurs in the calibration process.

F. Extrapolation Experiment

In this section, we will present the results from our extrapolation experiments on the duration QMF calibration approach. The extrapolation experiments were performed in order to test the robustness of the calibration approach to unseen values of the quality measures. There are two experiments in which we tested extrapolation performance of our proposed duration QMF calibration:

- *Short missing*: Calibration parameters were trained by using only 16 duration conditions by excluding durations of 5 s present in the model and/or test segments.
- *Long missing*: Calibration parameters were trained by using only 16 duration conditions by excluding conditions in which full condition present in the model and/or test segments.

These trained calibration parameters were then applied in evaluation to the nine duration conditions which were not seen in the calibration training. These extrapolation experiments

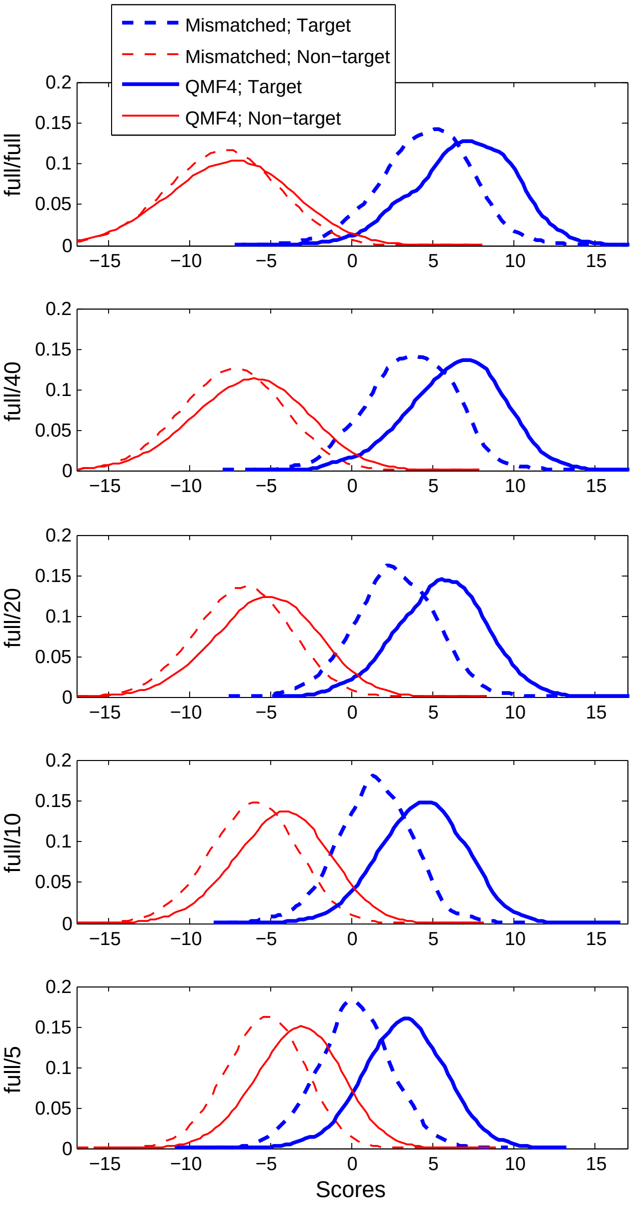
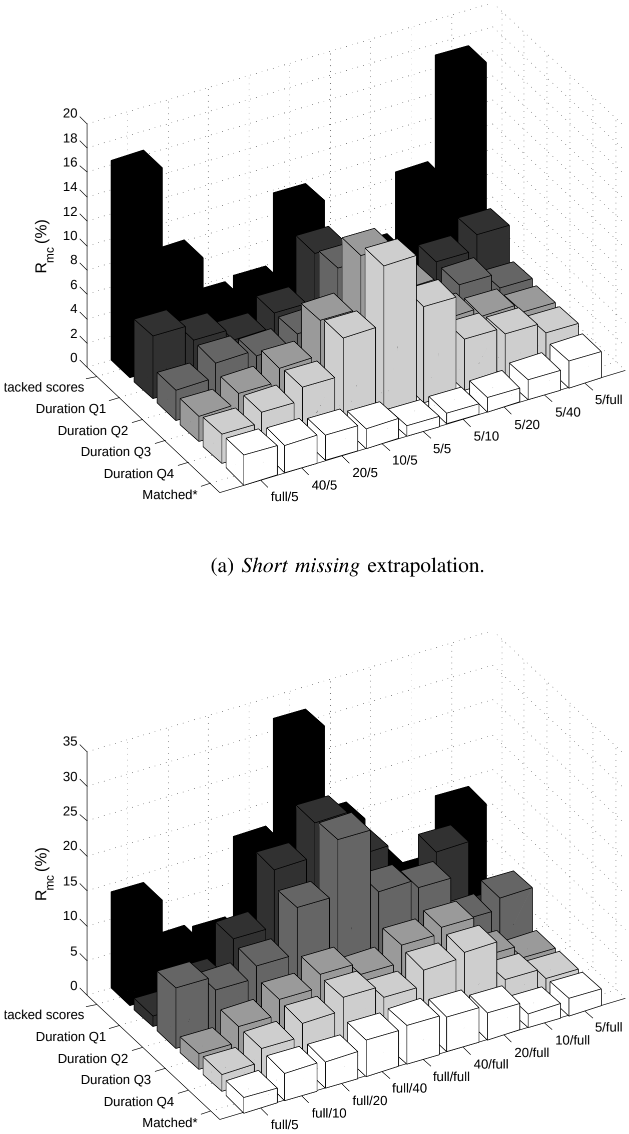


Fig. 5. Calibrated scores distribution for target and non-target trials from the mismatched (dashed line) and Q_4 (solid line) calibrations.

TABLE VIII
THE EXTRAPOLATION EXPERIMENT RESULTS IN THE SHORT MISSING AND LONG MISSING CONDITIONS FOR STACKED SCORES AND ALL DURATION QMFS CALIBRATION APPROACHES.

Calibration approach:	Short missing C_{llr}	C_{mc}	R_{mc} (%)	Long missing C_{llr}	C_{mc}	R_{mc} (%)
Mismatch*	.628	.147	34.09	.261	.076	28.98
Match*	.489	.008	1.67	.192	.006	3.78
Stacked scores	.519	.038	8.35	.207	.022	13.06
Q_1*	.493	.012	2.24	.196	.011	9.06
Q_1 Extrapolation	.504	.022	4.51	.197	.011	9.86
Q_2*	.495	.013	2.56	.195	.010	8.45
Q_2 Extrapolation	.503	.021	4.08	.204	.017	11.28
Q_3*	.496	.015	2.72	.193	.007	5.21
Q_3 Extrapolation	.509	.027	5.00	.194	.009	5.96
Q_4*	.497	.015	2.78	.193	.007	4.98
Q_4 Extrapolation	.510	.029	5.30	.194	.008	5.78

* No extrapolation experiments applied in this approach.



(a) *Short missing* extrapolation.

(b) *Long missing* extrapolation.

Fig. 6. Miscalibration rate values R_{mc} of extrapolation experiments on various calibration approaches in (a) *long missing* and (b) *short missing* conditions. Note that in the matched approach, extrapolation is not applicable.

were conducted for all four duration QMFs, and the results of these experiments are presented in Table VIII for both *short missing* and *long missing* extrapolations. The calibration metrics presented in Table VIII are averaged over the nine extrapolation conditions.

In order to be able to compare the performance of all calibration approaches in the extrapolation experiments, Figure 6 depicts the miscalibration rate values for every observed duration conditions in the short and long missing extrapolations. In both extrapolations, the system calibration performance drops compared to the experiments with calibration training on all 25 duration conditions. This may be expected because there is still an obvious duration mismatch. However, the extrapolated QMFs show a lot better performance than the stacked scores approach that did have access to the missing duration conditions.

In the *short missing* extrapolation, the Q_2 function has the best extrapolation performance compared to the other QMFs based on the average of miscalibration values which are presented in Table VIII. The Q_1 function has the next best performance, and both Q_1 and Q_2 perform statistically significantly better than the stacked calibration technique based on one-sided paired t-test similar to what we did for the results in Section V-E. As presented in Table VIII, the Q_3 and Q_4 functions do not perform as well as the other two in this extrapolation experiment, perhaps because they utilizes a too complex duration QMFs. Specifically in the short duration condition such as 5/5, Q_3 and Q_4 functions give limited performance improvement, similar to what we observed in the all duration training condition from the last paragraph of Section V-E.

In the *long missing* extrapolation, in contrary, the Q_3 and Q_4 functions offer statistically significantly better extrapolation performance compared to the Q_2 function, as we determined by a one-sided paired t-test. As it is mentioned in Section V-E, the Q_3 and Q_4 functions are, in fact, performing well in the presence of longer duration in both model and test segments. However, inspecting the difference of miscalibration average between the duration QMF approach for the full training and extrapolation experiment, the Q_1 and Q_2 functions show only a 0.001 and 0.003 absolute miscalibration increase, respectively. This is not very much and comparable to the increase found in Q_3 and Q_4 which shows the robustness of all proposed QMFs to unseen duration condition in calibration.

In general, the trend of the extrapolation results for both *short missing* and *long missing* for all duration QMFs indicate good extrapolation performance: in almost every evaluated condition in the extrapolation experiment, the calibration performance is better than the calibration performance of the stacked scores approach. We conclude that the proposed duration QMF approaches have successfully maintained a good calibration performance in the various duration conditions at the cost of adding at most two extra parameters in scores transformation for calibration, even if the durations have to be extrapolated beyond the range of durations seen in calibration training.

G. Comparison of Duration QMFs and k-means approaches

This section contrasts our proposed duration QMFs to other calibration techniques that also include quality measures. We use two contrasting approaches: calibration through *k-means clustering* of duration conditions and calibration via side-information using the BOSARIS toolkit [36] . Calibration through k-means clustering uses discrete calibration classes (clusters of duration) based on the duration in model and test segments. In applying calibration on the evaluation set, the parameters from the cluster closest to the evaluation train and test durations are used. The BOSARIS approach is the current common practice performing calibration for speaker recognition systems, and allows duration to be 'fused in' as side-information in the calibration step. In other parts of this paper, the calibration experiments were carried out using the FoCal toolkit, which is the predecessor of BOSARIS toolkit.

In order to evaluate the calibration approaches compared in this section, we generated a set of test segments with uniformly random durations in both model and test, ranging from 5 seconds to *full* length durations from the NIST SRE-2010 det-5 condition. The scores were then calibrated and evaluated using stacked scores, k-means clustering, BOSARIS, and QMFs approaches. In the k-means clustering approach, we use 25 clusters corresponding to the duration combinations used earlier, i.e., the calibration parameters were taken from the 'Matched Duration Condition'. In evaluation, the calibration parameters were used corresponding to the closest duration combination, e.g., a trial between a 12.5 s train segment and a 23.2 s test segment was calibrated with the $(10\,\text{s}, 20\,\text{s})$ calibration parameters from the Matched Duration Condition.

In the BOSARIS side-information, we used discrete classes for duration in both train and test as indicator vectors, e.g., for the example above the indicator vectors for train and test are $v_m = (0, 1, 0, 0, 0)^T$ and $v_t = (0, 0, 1, 0, 0)^T$ respectively. BOSARIS uses these in a bilinear fashion to train a *symmetric* offset matrix $\mathbf{V}$ using a term $v_m \mathbf{V} v_t^T$. As such, it is a symmetricized version of the k-means approach.

The calibrations results of these comparisons are presented in Table IX. As can be seen from the C_{mc} values, all calibration approaches produce very low miscalibration cost with k-means clustering has the highest cost. However, based on the $E_=$ and $C_{\text{llr}}^{\text{min}}$ values, the QMFs have generally better discrimination performance than other approaches. The Q_1 and Q_2 the best performance in both calibration and discrimination shown by all performance metrics presented in Table IX. With relatively less calibration parameters employed in the calibration process, the proposed duration QMF approach can outperform other techniques that also incur information of duration quality measures in calibration, e.g., with k-means clustering approach.

TABLE IX
CALIBRATION PERFORMANCE IN RANDOM TRUNCATED SCORES USING STACKED SCORES, K-MEANS CLUSTERING, BOSARIS, AND PROPOSED QMFS APPROACHES.

Calibration approach	C_{llr}	$C_{\text{llr}}^{\text{min}}$	C_{mc}	R_{mc} (%)	$E_=$ (%)	n_p
Stacked scores*	0.318	0.317	0.001	0.34	9.28	2
k-means clustering	0.312	0.311	0.002	0.48	9.19	50
BOSARIS	0.315	0.313	0.001	0.43	9.23	7
QMF 1	0.313	0.311	0.001	0.39	9.26	3
QMF 2	0.312	0.311	0.001	0.41	9.23	3
QMF 3	0.311	0.310	0.001	0.30	9.18	3
QMF 4	0.311	0.310	0.001	0.29	9.18	4

* Stacked scores does not include quality measures in calibration.

VI. CONCLUSION

Using a simple modification in the linear scores transformation for calibration by adding a quality measure function of duration is an easy and straight-forward idea to improve the calibration performance of speaker recognition system. This is observed from the calibration performance of the proposed duration quality measures approach and its comparison with other linear calibration approaches. Four duration quality measure functions are proposed and evaluated in this paper. All of them have their own advantages in counteracting the duration variability problem in calibration. Based on the one-sided paired t-test, all proposed QMFs perform statistically significantly better than the stacked scores calibration, and the saddle-shaped Q_3 and Q_4 functions offer better performance compared to the wedge-shaped Q_1 in terms of R_{mc}. We have also shown from the extrapolation experiments that the duration quality measures approach is fairly robust against the calibration problem of unseen duration condition in the calibration.

Future work in the topic of calibration with QMF technique includes using other quality measures such as background noise level which can be quantified as signal to noise ratio (SNR). Evaluation of the proposed duration QMFs are planned using different databases with more variation in duration conditions. With encouraging results achieved from the good calibration performance offered by QMF technique, further research on this topic is highly encouraged.

APPENDIX

In this appendix we will derive (9) following an argument put forward by Niko Brümmer. The relation is well-known by forensic statisticians, but we are not aware of any published derivation. The basic premise is that the likelihood ratio $\ell(x, y)$ for a speaker recognition system comparing speech samples x and y

$$\ell = \frac{P(x, y \mid H_1)}{P(x, y \mid H_2)} \tag{17}$$

is well-calibrated if it results in the same posterior distribution over H, whether ℓ or the speech input (x, y) is given. This means that all speaker comparison information is encoded in ℓ:

$$P(H|\ell) = P(H|x, y). \tag{18}$$

Applying Bayes' rule, and converting to the log odds domain this becomes

$$\log \frac{P(\ell|H_1)P(H_1)}{P(\ell|H_2)P(H_2)} = \log \frac{P(x, y|H_1)P(H_1)}{P(s, y|H_2)P(H_2)}, \tag{19}$$

where in the odds domain the factors with $P(\ell)$ and $P(x, y)$ cancel. In (19) the prior odds $P(H_1)/P(H_2)$ cancel as well, so that with the definition of the log likelihood ratio (17) we have

$$\log \frac{P(\ell|H_1)}{P(\ell|H_2)} = \log \frac{P(x, y|H_1)}{P(x, y|H_2)} = \ell, \tag{20}$$

which proves (9).

References

[1] R. J. Vogt, B. J. Baker, and S. Sridharan, "Modelling session variability in text independent speaker verification," in *International Speech Communication Association (ISCA)*.

[2] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Proc. of ICASSP*. IEEE, 2012, pp. 4249–4252.

[3] ——, "Evaluation of i-vector speaker recognition systems for forensic application," in *Proc. of Interspeech*, 2011, pp. 21–24.

[4] B. Fauve, N. Evans, N. Pearson, J. F. Bonastre, and J. Mason, "Influence of task duration in text-independent speaker verification," in *Proc. Interspeech*, vol. 7, 2007, pp. 794–797.

[5] A. Kanagasundaram, R. J. Vogt, D. B. Dean, and S. Sridharan, "PLDA based speaker recognition on short utterances," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*. ISCA, 2012.

[6] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1711–1723, 2007.

[7] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," *Proc. of Interspeech, Brisbane, Australia*, 2008.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[9] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.

[10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, 2001, pp. 213–218.

[11] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. of Interspeech*, 2006, pp. 1471–1474.

[12] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, 2010.

[13] M. McLaren and D. A. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Proc. of ICASSP*, 2011, pp. 5456–5459.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. of Interspeech*, 2011, pp. 249–252.

[15] P. M. Bousquet, A. Larcher, D. Matrouf, J. F. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, 2012.

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

[17] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Proc. Eurospeech*, September 2003, pp. 2017–2020.

[18] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration SVM-and GMM-based speaker verification," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, 2008.

[19] N. Brümmer, *FoCal-II: Toolkit for calibration of multi-class recognition scores*, August 2006, software available at http://www.dsp.sun.ac.za/∼nbrummer/focal/index.htm.

[20] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.

[21] N. Brümmer, "Spescom DataVoice NIST2005 SRE system description," in *Proc. NIST Speaker Recognition Evaluation Workshop*, 2004, toledo.

[22] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[23] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," *Speaker Classification I*, pp. 330–353, 2007.

[24] D. Ramos, "Forensic evaluation of the evidence using automatic speaker recognition systems," Ph.D. dissertation, Universidad Autonoma de Madrid, November 2007.

[25] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2104–2115, September 2007.

[26] C. S. Greenberg, "The NIST year 2012 speaker recognition evaluation plan," 2012. [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

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

[28] ——, "Using quality measures for multilevel speaker recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 192–209, 2006.

[29] L. Ferrer, K. Sönmez, and S. Kajarekar, "Class-dependent score combination for speaker recognition," in *Proc. Interspeech*, 2005.

[30] C. Chibelushi, F. Deravi, and J. Mason, "A review of speech-based bimodal recognition," *Multimedia, IEEE Transactions on*, vol. 4, no. 1, pp. 23–37, 2002.

[31] J. Kittler, N. Poh, O. Fatukasi, K. Messer, K. Kryszczuk, J. Richiardi, and A. Drygajlo, "Quality dependent fusion of intramodal and multimodal biometric experts," in *Proc. of SPIE Vol*, vol. 6539, 2007, pp. 653 903–1.

[32] J. Bigun, J. Fiérrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal biometric authentication using quality signals in mobile communications," in *Image Analysis and Processing, 2003. Proceedings. 12th International Conference on*. IEEE, 2003, pp. 2–11.

[33] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Discriminative multimodal biometric authentication based on quality measures," *Pattern Recognition*, vol. 38, no. 5, pp. 777–779, 2005.

[34] D. A. van Leeuwen, "The TNO SRE-2008 speaker recognition system," in *Proceedings of the NIST Speaker Recognition Evaluation Workshop*, Montreal, 2008.

[35] A. Strasheim and N. Brümmer, "Sunsdv system description: Nist sre 2008," NIST SRE workshop proceedings, May 2008.

[36] E. de Villiers and N. Brümmer, "Bosaris toolkit," 2010.

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

[38] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender independent speaker recognition," in *Proc. of Interspeech*, 2011, pp. 25–28.

[39] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *Proc. of ICASSP*. IEEE, 2011, pp. 4536–4539.

[40] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[41] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[42] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[43] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[44] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[45] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proc. of ICASSP*. IEEE, 2011, pp. 4832–4835.

[46] S. J. D. Prince, *Computer vision: models, learning, and inference*. Cambridge University Press, 2011.

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

[48] C. S. Greenberg, A. F. Martin, B. N. Barr, and G. R. Doddington, "Report on performance results in the NIST 2010 speaker recognition evaluation," in *Proc. of Interspeech*, 2011, pp. 261–264.

[49] National Institute of Standards and Technology, *NIST 2008 SRE Evaluation Plan*, available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/.

[50] ——, *NIST 2010 Speaker Recognition Evaluation Plan*, available at http://www.itl.nist.gov/iad/mig/tests/sre/2010/.

[51] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

[52] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch: University of Stellenbosch, 2010.

[53] J. Gonzalez-Rodriguez and D. Ramos, "Forensic automatic speaker classification in the coming paradigm shift," *Speaker Classification I*, pp. 205–217, 2007.

[54] M. Gebel, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Duisburg, University of Dortmund, 2009.

[55] D. A. van Leeuwen and N. Brümmer, "A speaker line-up for the likelihood ratio," in *Proc. Interspeech*. Firenze: ISCA, August 2011.

[56] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.

[57] D. A. van Leeuwen, *SRE Tools, Speaker Recognition Diagnostics in R*, available at https://sites.google.com/site/sretools/.

[58] T. Hasan, R. Saeidi, J. H. L. Hanson, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. ICASSP*. IEEE, 2013.

[59] J. A. Rice, *Mathematical statistics and data analysis*. Duxbury press, 2007.