# Modified-prior PLDA and Score Calibration for Duration Mismatch Compensation in Speaker Recognition System

*QingYang Hong*[1], *Lin Li*[1], *Ming Li*[2], *Ling Huang*[1], *Lihong Wan*[1], *Jun Zhang*[1]

[1]School of Information Science and Technology, Xiamen University, China
[2]SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University, China

lilin@xmu.edu.cn

## Abstract

To deal with the performance degradation of speaker recognition due to duration mismatch between enrollment and test utterances, a novel strategy to modify the standard normal prior distribution of the i-vector during probabilistic linear discriminant analysis (PLDA) modeling is employed. This new modified-prior PLDA model incorporates the covariance matrix scaled with duration of each utterance for each speaker, which achieves more discriminative characteristics by learning the duration variability as well as session variation in the i-vector space. Furthermore, an efficient Quality Measure Function (QMF) method which adopts duration variation as a compensation technique is employed to eliminate the linear shift in the score domain. To evaluate the robustness of the proposed approach, experiments were conducted on the NIST SRE10 core-core task in condition-5 with varying test utterance duration, in which the i-vectors of test utterances were extracted from full segment and randomly truncated segments of duration 10s and 20s. The results demonstrated the efficiency of modified-prior PLDA in different duration conditions, and the combined score calibration further improved the performance of speaker recognition.

**Index Terms**: speaker recognition, i-vector, duration mismatch, modified-prior PLDA, score calibration

## 1. Introduction

Automatic speaker recognition technology aims to distinguish the target speaker and the imposter by two main processing phases, feature extraction and model classification, which might be affected by several factors such as noise level, channel variation, utterance duration, speaker emotion, and so on. Recently, i-vector based speaker recognition systems [1] [2] have achieved the state-of-the-art performance on NIST evaluation tasks.

An i-vector is the low dimension representation of a Gaussian Mixture Model (GMM) mean supervector from a given speech utterance, which is obtained by evaluating the posterior expectation of the hidden variables on the Baum-Welch statistics from the Gaussian components of a Universal Background Model (UBM) [3] . Before 2012, the previous NIST evaluations controlled the noise level, utterance duration and speaker emotion, and most of the submissions assumed all the factors mentioned above degrading the performance of speaker recognition as a set of session uncertainty, which might be derived as the posterior covariance matrix to quantify the reliability of i-vector estimating process. Some session variability compensation techniques, including linear discriminant analysis (LDA), within-class covariance normalization (WCCN), nuisance attribute projection (NAP) and PLDA [4] are successfully used to model the speaker and channel subspace and attenuate the channel variability in i-vector speaker verification.

Usually, the duration of enrollment utterance is sufficient long, but the duration of test utterance may be very short. Under such condition of duration mismatch, the performance of the speaker recognition system using i-vector based PLDA degrades rapidly [5] . To figure out this mismatch problem, the relationship between i-vector length and duration variability was analyzed in [6] , and the experimental results demonstrated that i-vectors extracted from short utterances were less reliable than those from long utterances. Based on such analysis of i-vector length and variance, the quality metrics like duration in score calibration were further discussed to adjust the score distribution shifted by the duration mismatch [7] . Assuming duration variability as one of the major uncertainty associated with the i-vector extraction process, Kenny et al. [8] propagated this uncertainty in the PLDA model to obtain substantial improvements in accuracy at the cost of expensive computation during likelihood ratio maximization. Later, a short utterance variance modeling approach at i-vector level was introduced to compensate the session and duration variations and achieved improvement in performance [9] . Exploiting the duration influence as a posterior covariance in i-vector extraction, Sandro Cumani et al. [10] presented a new PLDA model with the intrinsic i-vector uncertainty and the proposed technique was declared to handle duration variability property with evaluation trials on NIST SRE 2010 and 2012.

We proposed an effective modified-prior PLDA framework in [11] to deal with the duration variation. As shorter utterances tend to have large covariance, the probability distribution function of i-vector can be modified with duration scaled covariance matrix during the PLDA training process. Then the formulation of the likelihood for standard Gaussian PLDA model is revised according to the duration-dependent posterior distribution of the i-vector. The results of evaluations on NIST SRE10 (condition-5) show that this modified-prior PLDA model outperforms the standard Gaussian PLDA when tested on variable duration.

Furthermore, this paper applied a QMF based calibration method as a compensation strategy of duration mismatch in the score domain. The evaluation metric $C_{llr}$ [12] and the relative loss ($R_{mc}$) were utilized to measure the validity of calibrated log-likelihood-ratios for a set of evaluation trials. The performance of the QMF based calibration on the proposed modified-prior PLDA system was evaluated in 9 calibration conditions, and the calibration experiments were further conducted in full duration version to analyze the performances of three recognition systems. The results showed that the duration quality measure approach applied on the modified-prior PLDA system was fairly robust against the mismatch duration problem.

## 2. Proposed method

In the state-of-the-art i-vector speaker recognition system, an i-vector ($x$) is a fixed-length vector extracted from the GMM mean supervector ($M$) based on the Baum-Welch statistics. Thus, every utterance is represented as an i-vector, which hypothesizes that both speaker and channel variabilities of speech utterances map into a single low dimensional subspace.

$$M = m + Tx \tag{1}$$

where $m$ is a speaker- and channel-independent supervector, $T$ is a rectangular matrix of low rank and $x$ is viewed as a random vector following a standard normal distribution N(0,I).

The value of the cosine kernel between the model i-vector $x_{mod}$ and the test i-vector $x_{tst}$ is adopted in this paper as a decision score of the baseline system.

$$s_{baseline} = \frac{< x_{tst}, x_{mod} >}{\|x_{tst}\| \, \|x_{mod}\|} \tag{2}$$

### 2.1. Standard Gaussian PLDA

PLDA has gained popularity as an elegant classification tool to find target classes in recent NIST challenges. In this paper, we use the Gaussian PLDA (G-PLDA) after i-vector length normalization [13] . Given a development set of i-vectors $x_{ij}$, $i = 1, ..., N, j = 1, ..., M_i$ from $N$ speakers ($M_i$ utterances for each speaker), each i-vector is distributed in a standard G-PLDA procedure as follows:

$$x_{ij} = \mu + \Phi\beta_i + \varepsilon_{ij} \tag{3}$$

where $\mu$ is the mean value generated from all i-vectors of the development set, $\beta_i$ is an identity variable of speaker $i$ having a standard normal prior $N(0, I)$, matrix $\Phi$ constrains the dimension of the speaker subspace, and the residual $\varepsilon_{ij}$ contains the session factors following a normal distribution with mean 0 and covariance matrix $\Sigma$ .

Supposed two i-vector $x_{mod}$ and $x_{tst}$ for model and test utterance respectively, the likelihood ratio between the same-speaker hypothesis $H_s$ and different-speaker hypothesis $H_d$ is formulated as[14] :

$$\begin{aligned} s(x_{mod}, x_{tst}) &= \log \frac{P(x_{mod}, x_{tst}|H_s)}{P(x_{mod}, x_{tst}|H_d)} \\ &= \log N\left(\begin{bmatrix} x_{mod} \\ x_{tst} \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma + \Phi\Phi^T & \Phi\Phi^T \\ \Phi\Phi^T & \Sigma + \Phi\Phi^T \end{bmatrix}\right) \\ &- \log N\left(\begin{bmatrix} x_{mod} \\ x_{tst} \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma + \Phi\Phi^T & 0 \\ 0 & \Sigma + \Phi\Phi^T \end{bmatrix}\right) \end{aligned} \tag{4}$$

### 2.2. Gaussian prior with duration factor

It is reported in [6] that the number of unique phones found in speech sample scales logarithmically with duration. For the same speaker, shorter segments tend to produce larger covariance, so that $\varepsilon_{ij}$ in (3) will follow a new normal distribution:

$$\varepsilon_{ij} \sim N(0, \Sigma) \rightarrow N(0, \Sigma_{eq,ij}) \tag{5}$$

where

$$\Sigma_{eq,ij} = \Sigma \cdot \left(\frac{L_{ij}}{\alpha}\right)^{-\lambda} \tag{6}$$

and $L_{ij}$ denotes the duration length of $j_{th}$ utterance for speaker $i$, $\alpha$ and $\lambda$ are the adjusting parameters which reflect the penalization degree on duration deviations. Typically, a value of $\alpha$ equals to mean duration of all i-vector and $\lambda$ can be set as a constant in the range of 0.5-2.0 based on actual experimental setup.

For $i = 1, ..., N, j = 1, ..., M_i$, let $\eta_{ij}$ denote the first order statistic $(x_{ij} - \mu)$, having the normal distribution $N(\Phi\beta_i, \Sigma_{eq,ij})$. Then the mean value of the first order statistic of speaker $i$ is defined as $F_i$,

$$F_i = \frac{\sum_{j=1}^{M_i} \eta_{ij}}{M_i} \tag{7}$$

and posterior distribution of $F_i$ is

$$P(F_i|\beta_i) = N\left(\Phi\beta_i, \frac{\Sigma_{eq,ij}}{M_i^2}\right) \tag{8}$$

Based on Bayes' theorem, we will get a relationship between the joint log-likelihood and PLDA parameters $(\Phi, \Sigma)$. With the EM algorithm, we estimate the parameters with termination condition when increment of the joint log-likelihood is less than the threshold $10^{-3}$ (the detailed derivation of formula was developed in [11]):

$$\begin{cases} \phi = \dfrac{\sum_{i=1}^{N}\left(\sum_{j=1}^{M_i}\left(\frac{L_{ij}}{a}\right)^{\lambda}\eta_{ij}\right)E(\beta_i)^T}{\sum_{i=1}^{N}\left(\sum_{j=1}^{M_i}\left(\frac{L_{ij}}{a}\right)^{\lambda}\right)E(\beta_i\beta_i^T)} \\ \Sigma = \dfrac{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left[\left(\frac{L_{ij}}{a}\right)^{\lambda}\eta_{ij}(\eta_{ij}^T - E(\beta_i)^T\phi^T)\right]}{\sum_{i=1}^{N} M_i} \end{cases} \tag{9}$$

Considering the duration scaled distribution of i-vector, the likelihood ratio of the proposed modified-prior PLDA system can be calculated in (4) using parameters $(\Phi, \Sigma_{eq,ij})$ instead of $(\Phi, \Sigma)$, which would increase the amount of calculation. Therefore, we use G-PLDA scoring method $s(\eta_{mod}, \eta_{tst})$ in experiments for computational simplification, and execute score calibration with duration variance for score compensation.

### 2.3. Score domain compensation

In this paper, a QMF based score calibration method [7] is adopted as a supplementary compensation to the proposed modified-prior PLDA to deal with the duration mismatch. Suppose that $\eta_{mod}$ and $\eta_{tst}$ are the first order statistics of i-vectors $x_{mod}$ and $x_{tst}$ respectively, then $s(\eta_{mod}, \eta_{tst})$ is defined as a trial score involving model utterance $x_{mod}$ and test utterance $x_{tst}$. The new score $q(\eta_{mod}, \eta_{tst})$ is obtained :

$$\begin{aligned} q(\eta_{mod}, \eta_{tst}) &= w_0 + w_1 s(\eta_{mod}, \eta_{tst}) \\ &+ w_2\big(\log\frac{d_{mod}}{d_c} + \frac{d_{tst}}{d_c}\big)^2 \\ &- w_3\big(\log\frac{d_{mod}}{d_c} - \frac{d_{tst}}{d_c}\big)^2 \end{aligned} \tag{10}$$

where $d_{mod}$ and $d_{tst}$ are related to duration of model segment and test segment respectively, and $d_c$ is a constant equal to the mean duration of all speech utterances.

The variances $w_0$ is the offset of the transformation and $w_1$ is a scaling parameter on original score $s(\eta_{mod}, \eta_{tst})$. Moreover, $w_2$ and $w_3$ are scaled to denote the duration conditions

Table 1: Evaluation performance of different systems on duration-mismatch trials.

| Duration | Baseline | | G-PLDA | | Proposed | |
|---|---|---|---|---|---|---|
| | $EER\%$ | minDCF | $EER\%$ | minDCF | $EER\%$ | minDCF |
| Full | 8.2 | 0.0309 | 4.1 | 0.0196 | 3.7 | 0.0178 |
| 20s | 12.7 | 0.0507 | 7.0 | 0.0320 | 6.5 | 0.0311 |
| 10s | 17.7 | 0.0608 | 10.4 | 0.0438 | 9.9 | 0.0464 |

Table 2: Score compensation of the proposed PLDA system evaluated on NIST SRE10 core-core task from referred scores of NIST SRE08 short2-short3 task of 9 calibration conditions. (Note: DRTN is the abbreviation of duration)

| Reference system | DRTN = Full | | | DRTN = 20s | | | DRTN = 10s | | |
|---|---|---|---|---|---|---|---|---|---|
| | $C_{llr}$ | $R_{mc}\%$ | $EER\%$ | $C_{llr}$ | $R_{mc}\%$ | $EER\%$ | $C_{llr}$ | $R_{mc}\%$ | $EER\%$ |
| DRTN = Full | 0.141 | 25.44 | 3.66 | 0.297 | 26.45 | 7.04 | 0.437 | 35.62 | 10.42 |
| DRTN = 20s | 0.149 | 32.86 | 3.57 | 0.262 | 11.22 | 6.75 | 0.611 | 89.77 | 9.86 |
| DRTN = 10s | 0.252 | 124.32 | 3.60 | 0.278 | 18.29 | 6.76 | 0.360 | 11.70 | 9.86 |

of model segment and test segment. All the four parameters can be obtained through the optimization on the basic measure $C_{llr}^{min}$ which minimums the cost of the log-likelihood-ratio $q(\eta_{mod}, \eta_{tst})$ using:

$$\begin{aligned}(w_1, w_2, w_3, w_4) &= \arg(C_{min}^{llr}) \\ &= \arg\min\{\frac{1}{2N_{tar}}\sum_{i\in tar}\log_2(1+\exp(-q_i)) \\ &+ \frac{1}{2N_{non}}\sum_{j\in non}\log_2(1+\exp(q_j))\}\end{aligned} \tag{11}$$

with $q_i$ and $q_j$ corresponding to the calibration score $q(\eta_{mod}, \eta_{tst})$ of target trial ($N_{tar}$ ) and non-target trial ($N_{non}$) respectively.

After the optimal transformation of scores that minimizes $C_{llr}$, the log-likelihood-ratio score is monotonously rising with the order of scores staying the same.

The performance of calibration can also be assessed on the miscalibration cost defined as the absolute loss ($C_{mc}$) and relative loss ($R_{mc}$):

$$C_{mc} = C_{llr} - C_{llr}^{min} \tag{12}$$

and

$$R_{mc} = \frac{C_{mc}}{C_{llr}^{min}} = \frac{C_{llr}}{C_{llr}^{min}} - 1 \tag{13}$$

# 3. Experiments

To analyze the effective performance of the proposed method to compensate duration mismatch, several evaluation tasks were designed on varying duration conditions with full version and randomly truncated test utterances with the duration of 10s and 20s respectively. All the experiments were conducted on NIST SRE10 core-core task (condition-5). We extracted 32-dimention MFCC with appended delta coefficients from each speech utterance. The total variability subspace of dimension 400 was estimated by the Baum-Welch statistics. And the PLDA was trained with speaker subspace of dimension 120. All the results presented in this paper concentrated on female trials only.

For NIST SRE10 core-core task(condition-5), 11370 utterances from NIST SRE04 and SRE05 corpora were picked out to train gender-dependent UBM containing 1024 Gaussians. And we used the same training data (24468 utterances) from NIST SRE04, SRE05, SRE06 and SRE08 corpora to estimate matrix T and PLDA parameters.

In the implementation of QMF based score calibration, the log-likelihood-ratio scores calculated from NIST SRE08 short2-short3 trials(condition-6) were used as the reference scores to calibrate the performance of trials on NIST SRE10. For NIST SRE08 trials, a gender-dependent UBM containing 512 Gaussians was trained on telephone speech data (about 1077 utterances) from SwitchBoard, NIST SRE04, SRE05 corpora. And total variability matrix T was estimated from 6063 utterances based on the same corpora as UBM training process. There were 10822 utterances from NIST SRE04, SRE05 and SRE06 corpora selected to train the parameters $(\Phi, \Sigma)$ of PLDA model.

### 3.1. Results of modified-prior PLDA

The modified-prior PLDA considering the duration as an important factor of covariance matrix was implemented to measure the discrimination performance of speaker recognition system on varying duration conditions. The equal error rate (EER) and the 2010 minimum decision cost function (minDCF) were calculated as evaluation metrics. As shown in Table 1, the proposed method outperformed the baseline system (CDS) and standard G-PLDA system for NIST SRE10 core-core task (condition-5). EER was reduced by 9.8% from 4.1% of G-PLDA to 3.7% of proposed method. And minDCF was reduced by 9.2% correspondingly. When the duration of test utterances decreased from full to 10s, the performance of all three systems became worse. However, the proposed modified-prior PLDA system suffered a most acceptable loss in terms of EER and minDCF.

### 3.2. Results of score compensation

The score compensation using QMF metric was evaluated on the SRE10 trials with parameters ($w_0$, $w_1$, $w_2$ and $w_3$) trained on SRE08 trials. We adopted the metric $C_{llr}^{min}$ as a representation of discrimination loss, and utilized miscalibration cost $R_{mc}$ to verify the calibration performance.

Experiments of the proposed duration QMF using different referred systems were implemented to train the calibration parameters ($w_0$, $w_1$, $w_2$ and $w_3$). Taking the implementations of modified-prior PLDA system on NIST SRE08 short2-short3

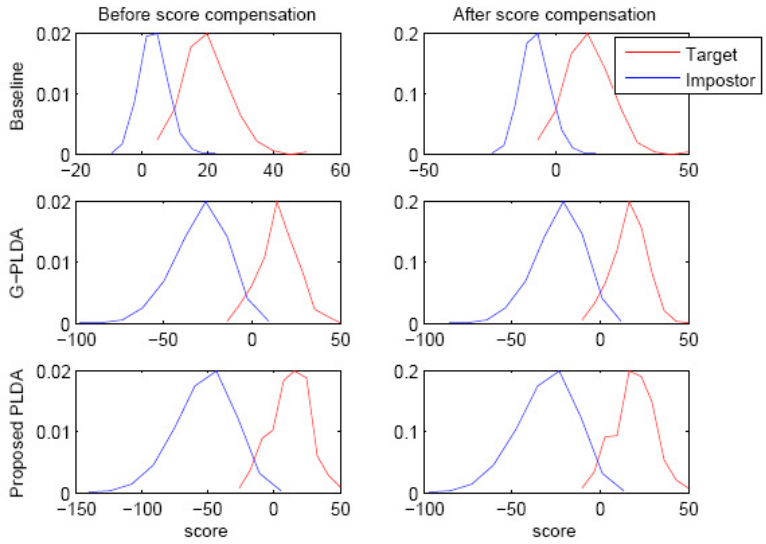


Figure 1: *Score distributions of NIST SRE10 (condition-5) database for before and after score compensation performed in three systems.*

Table 3: Score compensation of different systems evaluated on NIST SRE10 core-core task (full duration) from reference scores of NIST SRE08 short2-short3 task

| system (08 - 10) | $C_{llr}$ | $C_{llr}^{min}$ | $R_{mc}\%$ | $EER\%$ |
|---|---|---|---|---|
| Baseline | 0.301 | 0.2700 | 11.46 | 7.89 |
| G-PLDA | 0.153 | 0.1368 | 11.58 | 4.04 |
| Proposed | 0.141 | 0.1123 | 25.44 | 3.66 |

task (condition-6) as referred systems, the performance of the proposed PLDA system on NIST SRE10 was evaluated in 9 calibration conditions according to the test duration versions (full, 20s and 10s) as shown in Table 2. The results showed that $C_{llr}$, which was estimated from the calibration parameters of the same reference scores, increased gradually when test duration decreased from full duration to 10s. The minimum $C_{llr}$ was 0.141, which was achieved by the matched full duration test. And $C_{llr}$ of matched 20s and 10s were 0.262 and 0.360 respectively. Moreover, the least calibration loss was achieved by QMF technique under the matched duration condition according to $R_{mc}$ value.

To further analyze the effect of calibration on score domain, experiments of QMF-based score calibration were performed in full duration version by three recognition systems: baseline, standard G-PLDA and modified-prior PLDA. As can be seen from Figure 1, the distributions of scores from baseline system, standard G-PLDA system and the proposed PLDA system were skewed to the assumptions of application-independent decision of target speaker or impostor. After score calibration, all the scores distributions of three recognition systems were normalized around the center of score zero, which would behave more robust and efficient. In Table 3, we could also observe that good performance was obtained in both calibration and discrimination for the proposed scheme utilizing the quality measure function with modified-prior PLDA. EER was reduced by 9.4% from 4.04% of G-PLDA to 3.66% of proposed method. Especially, the best optimization of $C_{llr}$ was achieved by calibration on the modified-prior PLDA system. However, the value of $R_{mc}\%$ indicated that further research on QMF based calibration might be encouraged.

## 4. Conclusions

Duration is one of the major mismatch factors in speaker recognition system. To discriminate the personality characteristics of each utterance and decrease the influence of duration, a new strategy exploiting duration as a scaling parameter in standard G-PLDA procedure was proposed in this paper. Without the prior knowledge of i-vector extractor, the proposed PLDA model revised the distribution of each i-vector for each speaker only regarding the intrinsic length of each utterance. Furthermore, the proposed modified-prior PLDA method combining the duration-based QMF score calibration performed significantly better than the systems using only duration optimization.

## 5. Acknowledgment

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61105026 and No. 11274259).

## 6. References

[1] N. Brummer, L. Burget, P. Kenny, P. Matejka, E. de Villiers, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, D. Baum *et al.*, "Abc system description for nist sre 2010," *Proc. NIST 2010 Speaker Recognition Evaluation*, pp. 1–20, 2010.

[2] N. Dehak, P. 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.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[4] S. J. 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.

[5] A. K. Sarkar, D. Matrouf, P.-M. Bousquet, and J.-F. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification." in *INTERSPEECH*, 2012.

[6] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recogni-

tion systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7663–7667.

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

[8] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7649–7653.

[9] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication*, vol. 59, pp. 69–82, 2014.

[10] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in probabilistic linear discriminant analysis," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 846–857, 2014.

[11] C. Weicheng, L. Ming, L. Lin, and H. Qingyang, "Duration dependent covariance regularization in plda modeling for speaker verification," *submitted to Interspeech*, 2015.

[12] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," *arXiv preprint arXiv:1311.0707*, 2013.

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

[14] S. Cumani, N. Brummer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the-vector space," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 1217–1227, 2013.