POLITECNICO DI TORINO
Repository ISTITUZIONALE

A Generative Model for Duration-Dependent Score Calibration

Original
A Generative Model for Duration-Dependent Score Calibration / Cumani, Sandro; Sarni, Salvatore. - ELETTRONICO. - (2021), pp. 4598-4602. (Interspeech 2021 Brno (Repubblica Ceca) 30 August - 3 September 2021) [10.21437/Interspeech.2021-114].

Availability:
This version is available at: 11583/2922679 since: 2021-09-09T15:30:49Z

Publisher:
ISCA

Published
DOI:10.21437/Interspeech.2021-114

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

25 April 2026

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A generative model for duration-dependent score calibration

Sandro Cumani, Salvatore Sarni

Politecnico di Torino, Italy
sandro.cumani@polito.it, salvatore.sarni@polito.it

Abstract

In this work we introduce a generative score calibration model for speaker verification systems able to explicitly account for utterance-dependent miscalibration sources, with a focus on segment duration. The model is theoretically motivated by an analysis of the effects of distribution mismatch on the scores produced by Probabilistic Linear Discriminant Analysis (PLDA), and extends our previous investigation on the distribution of well-calibrated PLDA log-likelihood ratios. We characterize target and non-target scores by means of Variance-Gamma densities, whose parameters represent effective between and within-class variabilities. Experimental results on SRE 2019 show that the proposed method improves both calibration and verification accuracy with respect to duration-agnostic models and to duration-aware discriminative methods.
Index Terms: Generative score calibration, Variance-Gamma distribution, domain mismatch, duration-dependent calibration

1. Introduction

The standard approach for calibration of speaker verification scores is based on discriminative prior-weighted Logistic Regression (Log-Reg) [1, 2], successfully employed in a plethora of different scenarios [3, 4, 5, 6, 7]. Logistic regression is effective in reducing miscalibration, and allows incorporating side-information, the most important being segment duration, to improve not only calibration, but also the accuracy of speaker verification systems [8, 9, 10]. Generative models have recently proven to be a viable alternative to discriminative methods that allows both for supervised and unsupervised training. In [11] the authors analyzed the constraints that well-calibrated score distributions should satisfy, and proposed a simple yet effective linear calibration model based on constrained Gaussian distributions. The model was further extended in [12] to handle missing labels. In [13] the authors propose to model target and non-target scores with different, unconstrained densities, including T–student and Normal Inverse Gaussians (NIG).

Recently, we have proposed a generative linear model based on Variance-Gamma (VΓ) distributions [14, 15, 16]. The model was motivated by our analysis of the distribution of well-calibrated log-likelihood ratios (LLR) obtained by Probabilistic Linear Discriminant Analysis [17, 18] (PLDA) classifiers, and has proven to be effective not only for supervised tasks, where it matched linear logistic regression, but also for unsupervised scenarios, where it outperformed other generative approaches. This method, however, has two main limitations. Since the calibration model is linear, it may not be effective in presence of non-linear miscalibration effects. For supervised scenarios non-linear approaches such as the unconstrained NIG method of [13] can provide better calibration, although we have shown in [14] that the additional freedom of unconstrained models can be detrimental for unsupervised tasks. Furthermore, current state-of-the-art generative models are not able to effectively account for trial-dependent miscalibration sources such as utterance duration. In this work we address these limitations by investigating the distribution of scores of PLDA classifiers in presence of speaker vector distribution mismatch. In Sections 2 and 3 we show how the between and within-class covariances of the evaluation population affect the distribution of scores of Gaussian-distributed speaker vectors. This allows us to introduce, in Section 4, a non-linear Variance-Gamma score model, whose parameters represent effective between and within-class variability and can account for trial-level mismatch sources. As we show in Sections 5 and 6, a simple duration model can be combined with the score model to achieve state-of-the-art calibration for utterances of variable duration. This is, to our knowledge, the first successful attempt at incorporating duration effects in generative calibration approaches.

2. The distribution of well-calibrated PLDA scores

We consider the simplified two-covariance PLDA model[1]

$$\boldsymbol{\Phi} = \overline{\mathbf{Y}} + \overline{\mathbf{E}}\,, \tag{1}$$

where $\boldsymbol{\Phi}$ is the M-dimensional Random Variable (R.V.) responsible for generating an observed speaker vector $\boldsymbol{\phi}$ (e.g. i-vectors [20], e-vectors [21], or speaker embeddings [22]), $\overline{\mathbf{Y}}$ is the R.V. representing the speaker identity and $\overline{\mathbf{E}}$ represents residual noise. The prior distributions of $\overline{\mathbf{Y}}$ and $\overline{\mathbf{E}}$ are:

$$\overline{\mathbf{Y}} \sim \mathcal{N}(\mathbf{m}_{\mathcal{M}}, \mathbf{B}_{\mathcal{M}})\,, \qquad \overline{\mathbf{E}} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_{\mathcal{M}})\,. \tag{2}$$

Without loss of generality, we assume that $\mathbf{m}_{\mathcal{M}} = 0$, and both $\mathbf{B}_{\mathcal{M}}$ and $\mathbf{W}_{\mathcal{M}}$ are diagonal. Let $\mathbf{z} = \left[\boldsymbol{\phi}_{\mathcal{E}}^T, \boldsymbol{\phi}_{\mathcal{T}}^T\right]^T$ be a pair of speaker vectors (enroll and test), realization of R.V. $\mathbf{Z} = \left[\boldsymbol{\Phi}_{\mathcal{E}}^T, \boldsymbol{\Phi}_{\mathcal{T}}^T\right]^T$. According to (1), under the same and different speaker hypotheses $\mathfrak{S}$ and $\mathfrak{D}$ the distribution of $\mathbf{Z}$ is

$$\begin{bmatrix}\boldsymbol{\Phi}_{\mathcal{E}} \\ \boldsymbol{\Phi}_{\mathcal{T}}\end{bmatrix} | \mathfrak{S} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{S}}\right) \quad \begin{bmatrix}\boldsymbol{\Phi}_{\mathcal{E}} \\ \boldsymbol{\Phi}_{\mathcal{T}}\end{bmatrix} | \mathfrak{D} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{D}}\right) \tag{3}$$

with

$$\boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{S}} = \begin{bmatrix}\mathbf{T}_{\mathcal{M}} & \mathbf{B}_{\mathcal{M}} \\ \mathbf{B}_{\mathcal{M}} & \mathbf{T}_{\mathcal{M}}\end{bmatrix}\,, \quad \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{D}} = \begin{bmatrix}\mathbf{T}_{\mathcal{M}} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{\mathcal{M}}\end{bmatrix}\,, \tag{4}$$

and $\mathbf{T}_{\mathcal{M}} = \mathbf{B}_{\mathcal{M}} + \mathbf{W}_{\mathcal{M}}$. The log-likelihood ratio for a pair of speaker vectors $\mathbf{z}$ is

$$\ell(\mathbf{z}) = K_{\mathcal{M}} - \frac{1}{2}\mathbf{z}^T\left(\boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{S}}^{-1} - \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{D}}^{-1}\right)\mathbf{z}\,. \tag{5}$$

where $K_{\mathcal{M}} = \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{D}}\right|$. As we showed in [14], if an evaluation trial is a realization of $\mathbf{Z}$, its score can be interpreted as a realization of R.V. $\mathcal{L}$, conditionally defined as

$$\mathcal{L}|\mathfrak{D} = \ell(\mathbf{Z}|\mathfrak{D})\,, \quad \mathcal{L}|\mathfrak{S} = \ell(\mathbf{Z}|\mathfrak{S})\,.$$

[1] Extension to subspace-constrained models is straightforward — see, for example, [19] for the relationship between the two models.

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-114

$\mathcal{L}|\mathfrak{S}$ and $\mathcal{L}|\mathfrak{D}$ can be expressed as sums of M independent Variance-Gamma distributed R.V.s whose parameters depend on the between-to-within variability ratios $\rho_i = \frac{\mathbf{B}_{\mathcal{M},i}}{\mathbf{W}_{\mathcal{M},i}}$, where $\mathbf{B}_{\mathcal{M},i}$ and $\mathbf{W}_{\mathcal{M},i}$ are the i-th elements of the diagonals of $\mathbf{B}_{\mathcal{M}}$ and $\mathbf{W}_{\mathcal{M}}$, respectively. If we also assume isotropic $\rho_i \triangleq \rho$, a score can be interpreted as a realization of R.V. $\mathcal{L}$, with

$$\mathcal{L}|\mathfrak{D} \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta_{\mathfrak{D}}, \mu)\ , \quad \mathcal{L}|\mathfrak{S} \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta_{\mathfrak{S}}, \mu)\ , \quad (6)$$

$$\beta_{\mathfrak{D}} = -1\ , \quad \alpha^2 = \frac{(\rho+1)^2}{\rho^2}\ , \quad \beta_{\mathfrak{S}} = 0\ , \quad \lambda = \frac{M}{2}\ .$$

In [14] we assumed that most miscalibration effects can be captured through a linear miscalibration model, where observed scores are obtained as an affine transformation of well-calibrated, VΓ-distributed scores. Since the transformation alone is not able to account for the skewness that is often observed in empirical score distributions, we relaxed the assumptions on $\beta_{\mathfrak{S}}$ and $\beta_{\mathfrak{D}}$, allowing for non-zero values for $\beta_{\mathfrak{S}}$, and tying the two parameters as $\beta_{\mathfrak{S}} = \beta_{\mathfrak{D}} + 1$.

3. The distribution of PLDA scores in presence of domain mismatch

Well-calibrated VΓ densities model score distributions in which evaluation trials are sampled according to the PLDA model (3). In most cases, however, the mismatch between the PLDA model and the evaluation population is significant (if this was not the case, score calibration would not be required). To investigate how this affects the distribution of PLDA scores, we assume that scores are computed using the PLDA model parameters in (5), but evaluation trials are not samples of R.V.s distributed as in (3), but rather of R.V. $\mathbf{Z} = [\mathbf{\Phi}_{\mathcal{E}}, \mathbf{\Phi}_{\mathcal{T}}]$ with distribution

$$\begin{bmatrix}\mathbf{\Phi}_{\mathcal{E}}\\ \mathbf{\Phi}_{\mathcal{T}}\end{bmatrix}|\mathfrak{S} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma}_{\mathfrak{S}})\ , \quad \begin{bmatrix}\mathbf{\Phi}_{\mathcal{E}}\\ \mathbf{\Phi}_{\mathcal{T}}\end{bmatrix}|\mathfrak{D} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma}_{\mathfrak{D}})\ , \quad (7)$$

where

$$\mathbf{\Sigma}_{\mathfrak{S}} = \begin{bmatrix}\mathbf{T}_{\mathcal{E}} & \mathbf{B}_{\mathcal{C}}\\ \mathbf{B}_{\mathcal{C}} & \mathbf{T}_{\mathcal{T}}\end{bmatrix}\ , \quad \mathbf{\Sigma}_{\mathfrak{D}} = \begin{bmatrix}\mathbf{T}_{\mathcal{E}} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}_{\mathcal{T}}\end{bmatrix}\ , \quad (8)$$

$\mathbf{T}_{\mathcal{E}} = \mathbf{B}_{\mathcal{C}} + \mathbf{W}_{\mathcal{E}}$ and $\mathbf{T}_{\mathcal{T}} = \mathbf{B}_{\mathcal{C}} + \mathbf{W}_{\mathcal{T}}$. Matrix $\mathbf{B}_{\mathcal{C}}$ represents the common between-class variability for the evaluation population, whereas $\mathbf{W}_{\mathcal{E}}$ and $\mathbf{W}_{\mathcal{T}}$ represent within-class variability for the enrollment and test segments. We directly consider the more general case with different enroll and test within-class covariance matrices as this will lead, in Section 5, to a simple solution for modeling utterance duration. To keep the model tractable, we also assume that $\mathbf{B}_{\mathcal{C}}$, $\mathbf{W}_{\mathcal{E}}$ and $\mathbf{W}_{\mathcal{T}}$ are diagonal. We will show that, despite this assumption, the resulting model is powerful enough to improve calibration with respect to the linear VΓ model of [14]. We also assume[2] $\mathbf{m} = \mathbf{0}$. Given that all covariance matrices are diagonal, we can again represent the LLR as a sum of M independent terms

$$\ell(\mathbf{z}) = \sum_{i=1}^{M} \frac{1}{2}\mathbf{z}_i^T \mathbf{A}_i \mathbf{z}_i + k_i\ , \quad (9)$$

where $\mathbf{z}_i = \left[\boldsymbol{\phi}_{\mathcal{E},i}\ , \boldsymbol{\phi}_{\mathcal{T},i}\right]^T$ stacks the i-th components of the speaker vectors $\boldsymbol{\phi}_{\mathcal{E}}, \boldsymbol{\phi}_{\mathcal{T}}$, and

$$\mathbf{A}_i = \frac{\mathbf{B}_{\mathcal{M},i}}{\mathbf{T}^2_{\mathcal{M}_i} - \mathbf{B}^2_{\mathcal{M},i}} \begin{bmatrix} -\frac{\mathbf{B}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{M}_i}} & 1 \\ 1 & -\frac{\mathbf{B}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{M}_i}} \end{bmatrix}\ , \quad (10)$$

[2]The evaluation population mean can be easily estimated and compensated from few, unlabeled evaluation samples. In Section 4 we will show a possible way to also account for small, non-zero $\mathbf{m}$.

with $\mathbf{T}_{\mathcal{M}_i} = \mathbf{B}_{\mathcal{M},i} + \mathbf{W}_{\mathcal{M},i}$. The terms k_i correspond to $k_i = \frac{1}{2}\log \mathbf{T}^2_{\mathcal{M},i} - \frac{1}{2}\log\left(\mathbf{T}^2_{\mathcal{M},i} - \mathbf{B}^2_{\mathcal{M},i}\right)$. Since we assumed that $\mathbf{B}_{\mathcal{E}}$,$\mathbf{W}_{\mathcal{E}}$ and $\mathbf{W}_{\mathcal{T}}$ are diagonal, the R.V.s corresponding to the i-th components of the speaker vectors $\mathbf{Z}_i = [\mathbf{\Phi}_{\mathcal{E},i}\ , \mathbf{\Phi}_{\mathcal{T},i}]$ are independent under both the same and different speaker hypotheses, thus we can write $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ as sums of M independent R.V.s as:

$$\mathcal{L}|\mathfrak{D} = \ell(\mathbf{Z}|\mathfrak{D}) = \sum_{i=1}^{M} \mathcal{L}_i|\mathfrak{D}\ , \quad \mathcal{L}|\mathfrak{S} = \ell(\mathbf{Z}|\mathfrak{S}) = \sum_{i=1}^{M} \mathcal{L}_i|\mathfrak{S}\ , \quad (11)$$

where

$$\mathcal{L}_i|\mathfrak{h} = \ell(\mathbf{Z}_i|\mathfrak{h}) = \frac{1}{2}\mathbf{Z}^T_{\mathfrak{h},i}\mathbf{A}_i\mathbf{Z}_{\mathfrak{h},i} + k_i\ , \quad (12)$$

with $\mathfrak{h} \in \{\mathfrak{S}, \mathfrak{D}\}$, $\mathbf{Z}_{\mathfrak{h},i} \sim \mathbf{Z}_i|\mathfrak{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathfrak{h},i})$, and

$$\mathbf{\Sigma}_{\mathfrak{S},i} = \begin{bmatrix}\mathbf{T}_{\mathcal{E},i} & \mathbf{B}_{\mathcal{C},i}\\ \mathbf{B}_{\mathcal{C},i} & \mathbf{T}_{\mathcal{T},i}\end{bmatrix}\ , \quad \mathbf{\Sigma}_{\mathfrak{D},i} = \begin{bmatrix}\mathbf{T}_{\mathcal{E},i} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}_{\mathcal{T},i}\end{bmatrix}\ . \quad (13)$$

$\mathbf{B}_{\mathcal{C},i}$, $\mathbf{T}_{\mathcal{E},i}$ and $\mathbf{T}_{\mathcal{T},i}$ are the i-th elements of the diagonals of $\mathbf{B}_{\mathcal{C}}$, $\mathbf{T}_{\mathcal{E}}$ and $\mathbf{T}_{\mathcal{T}}$, respectively. Let $\mathbf{C}_{\mathfrak{h},i}$ denote the Cholesky decomposition of $\mathbf{\Sigma}_{\mathfrak{h},i}$. Let also $\mathbf{U}_{\mathfrak{h},i}\mathbf{D}_{\mathfrak{h},i}\mathbf{U}^T_{\mathfrak{h},i}$ denote the eigendecomposition of $\mathbf{M}_{\mathfrak{h},i} \triangleq \mathbf{C}^T_{\mathfrak{h},i}\mathbf{A}_i\mathbf{C}_{\mathfrak{h},i} = \mathbf{U}_{\mathfrak{h},i}\mathbf{D}_{\mathfrak{h},i}\mathbf{U}^T_{\mathfrak{h},i}$. The conditional distributions of $\mathcal{L}_i$ can then be rewritten as

$$\mathcal{L}_i|\mathfrak{h} \sim \frac{1}{2}\mathbf{Y}^T\mathbf{C}^T_{\mathfrak{h},i}\mathbf{A}\mathbf{C}_{\mathfrak{h},i}\mathbf{Y} + k_i \sim \frac{1}{2}\mathbf{Y}^T\mathbf{D}_{\mathfrak{h},i}\mathbf{Y} + k_i\ , \quad (14)$$

where $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard normal distributed R.V., and $\mathbf{D}_{\mathfrak{h},i}$ contains the eigenvalues of $\mathbf{M}_{\mathfrak{h},i}$. The determinant of $\mathbf{M}_{\mathfrak{h},i}$ is negative, therefore its two eigenvalues have different sign. To derive an expression for $\mathcal{L}_i$ we analyze the distribution of quadratic, indefinite forms (14). For the sake of readability, we drop all suffices and we simply consider quadratic forms

$$\mathcal{L} = \frac{1}{2}\mathbf{Y}^T\mathbf{D}\mathbf{Y} + k = \frac{1}{2}d_+Y_+^2 - \frac{1}{2}|d_-|Y_-^2 + k \quad (15)$$

where $\mathbf{D}$ is a 2×2 diagonal matrix with elements $d_+ > 0$ and $d_- < 0$, and $\mathbf{Y} = [Y_+, Y_-]^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We denote

$$G_+ \triangleq \frac{1}{2}d_+Y_+^2\ , \quad G_- \triangleq \frac{1}{2}|d_-|Y_-^2\ , \quad \mathcal{L} = G_+ - G_- + k\ . \quad (16)$$

Since Y_+ and Y_- are independent and standard normal distributed, $Y_+ \sim Y_- \sim \mathcal{N}(0,1)$, G_+ and G_- are also independent, and Gamma distributed:

$$G_+ \sim \Gamma\left(\frac{1}{2}, \frac{1}{d_+}\right)\ , \quad G_- \sim \Gamma\left(\frac{1}{2}, \frac{1}{|d_-|}\right)\ , \quad (17)$$

and therefore $\mathcal{L}$ follows a Variance-Gamma distribution [23, 14]. To derive the parameters, we consider the Moment Generating Function (MGF) of $\mathcal{L}$:

$$\begin{aligned} M_{\mathcal{L}}(t) &= e^{kt}M_{G_+}(t)M_{G_-}(-t) \\ &= e^{kt}\left(1 - \mathrm{tr}(\mathbf{D})t + \det(\mathbf{D})t^2\right)^{-\frac{1}{2}}\ , \end{aligned} \quad (18)$$

where tr and det denote the trace and determinant operators, respectively. The MGF of a VΓ distributed R.V. $X \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta, \mu)$ is

$$M_X(t) = e^{\mu t}\left(1 - \frac{2\beta}{\gamma^2}t - \frac{t^2}{\gamma^2}\right)^{-\lambda}\ , \quad (19)$$

with $\gamma^2 = \alpha^2 - \beta^2$. We can verify that $\mathcal{L}$ is VΓ-distributed, and the parameters can be recovered by inspection:

$$\lambda = \frac{1}{2}\,, \quad \mu = k\,, \quad \gamma^2 = -\frac{1}{\det(\mathbf{D})}\,, \quad \beta = -\frac{1}{2}\frac{\mathrm{tr}(\mathbf{D})}{\det(\mathbf{D})}\,. \tag{20}$$

Since $\mathrm{tr}(\mathbf{D}_{\mathfrak{h},i}) = \mathrm{tr}(\mathbf{M}_{\mathfrak{h},i}) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h},i})$ and $\det(\mathbf{D}_{\mathfrak{h},i}) = \det(\mathbf{M}_{\mathfrak{h},i}) = \det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h},i})$, $\mathcal{L}_i|\mathfrak{D}$ and $\mathcal{L}_i|\mathfrak{S}$ are VΓ-distributed:

$$\mathcal{L}_i|\mathfrak{h} \sim \mathrm{V\Gamma}\left(\frac{1}{2}, k_i, \alpha_{\mathfrak{h},i}, \beta_{\mathfrak{h},i}\right)\,, \tag{21}$$

with $\alpha^2_{\mathfrak{h},i} = \gamma^2_{\mathfrak{h},i} + \beta^2_{\mathfrak{h},i}$, and

$$\beta_{\mathfrak{h},i} = -\frac{1}{2}\frac{\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h},i})}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h},i})}\,,\ \gamma^2_{\mathfrak{h},i} = -\frac{1}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h},i})}\,. \tag{22}$$

In many cases it's reasonable to assume that the enroll and test speaker vectors are affected by independent, identically distributed (i.i.d.) noise, distributed as $\mathcal{N}(\mathbf{0}, \mathbf{W}_{\mathcal{C}})$, so that $\mathbf{W}_{\mathcal{E}} = \mathbf{W}_{\mathcal{T}} = \mathbf{W}_{\mathcal{C}}$. After some algebraic manipulations we can write the parameters of the score distributions:

$$\gamma^2_{\mathfrak{D},i} = \frac{\mathbf{T}^2_{\mathcal{M},i}}{\mathbf{T}^2_{\mathcal{C},i}}\frac{1+2\rho_{\mathcal{M},i}}{\rho^2_{\mathcal{M},i}}\,, \quad \gamma^2_{\mathfrak{S},i} = \gamma^2_{\mathfrak{D},i}\frac{(1+\rho_{\mathcal{C},i})^2}{1+2\rho_{\mathcal{C},i}} \tag{23}$$

$$\beta_{\mathfrak{D},i} = -\frac{\mathbf{T}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{C},i}}\,, \quad \beta_{\mathfrak{S},i} = \frac{\mathbf{T}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{C},i}}\frac{1+\rho_{\mathcal{C},i}}{1+2\rho_{\mathcal{C},i}}\left(\frac{\rho_{\mathcal{C},i}}{\rho_{\mathcal{M},i}} - 1\right),$$

where $\mathbf{T}_{\mathcal{C}} = \mathbf{W}_{\mathcal{C}} + \mathbf{B}_{\mathcal{C}}$. We can observe that the model parameters depend only on three values: the ratio $\frac{\mathbf{T}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{C},i}}$, representing the scale of test samples compared to the scale of training samples, and the between-to-within variability ratios of the training and evaluation populations $\rho_{\mathcal{M},i}$ and $\rho_{\mathcal{C},i}$. Furthermore, in contrast with well-calibrated LLRs, in this case both $\mathcal{L}_i|\mathfrak{S}$ and $\mathcal{L}_i|\mathfrak{D}$ may be skewed. In particular, the non-target distribution is left-skewed, whereas the skewness of the target distribution depends on whether $\rho_{\mathcal{C},i}$ is greater or smaller than $\rho_{\mathcal{M},i}$: in the former case, the distribution will be right-skewed (this corresponds to evaluation vectors that are easier to discriminate with respect to training samples along direction i), in the latter case the distribution will be left-skewed. Finally, we can observe that the non-target distribution does not depend on the evaluation population between-over-within variability ratio $\rho_{\mathcal{C},i}$.

4. A generative model for mismatched data

In general, the distribution of $\mathcal{L}|\mathfrak{S}$ and $\mathcal{L}|\mathfrak{D}$ cannot be expressed in closed form. However, if we assume that, for all directions i,

$$\begin{aligned}&\mathbf{B}_{\mathcal{M},i} = \xi_i b_{\mathcal{M}}\,, \mathbf{W}_{\mathcal{M},i} = \xi_i w_{\mathcal{M}}\\ &\mathbf{B}_{\mathcal{C},i} = \xi_i b_{\mathcal{C}}\,, \quad \mathbf{W}_{\mathcal{E},i} = \xi_i w_{\mathcal{E}}\,,\ \mathbf{W}_{\mathcal{T},i} = \xi_i w_{\mathcal{T}}\,,\end{aligned} \tag{24}$$

for scalars $\xi_i \neq 0$, then $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ are VΓ-distributed as

$$\mathcal{L}|\mathfrak{h} \sim \mathrm{V\Gamma}\left(\frac{M}{2}, \sum_{i=1}^{M} k_i, \alpha_{\mathfrak{h}}, \beta_{\mathfrak{h}}\right)\,, \tag{25}$$

where $\alpha_{\mathfrak{h}}$ and $\beta_{\mathfrak{h}}$ can be computed from (10) and (22) using the parameters for any of the i-th directions, since the result does not depend on the values ξ_i. Assuming isotropic normalized variances is a strong approximation which, however, is effective as long as we also estimate a "effective" number of speaker vector dimensions through the parameter λ of the VΓ distributions [14]. Therefore, we consider the score model

$$\mathcal{L}|\mathfrak{h} \sim \mathrm{V\Gamma}\left(\lambda, \mu_{\mathfrak{h}}, \alpha_{\mathfrak{h}}, \beta_{\mathfrak{h}}\right)\,, \tag{26}$$

where λ is a shared shape parameter, $\mu_{\mathfrak{h}}$ are location parameters and $\alpha_{\mathfrak{h}}, \beta_{\mathfrak{h}}$ depend on the parameters $b_{\mathcal{M}}, w_{\mathcal{M}}, b_{\mathcal{C}}, w_{\mathcal{E}}, w_{\mathcal{T}}$ through

$$\begin{aligned}&t_{\mathcal{M}} = b_{\mathcal{M}} + w_{\mathcal{M}}\,, \quad t_{\mathcal{E}} = b_{\mathcal{C}} + w_{\mathcal{E}}\,, \quad t_{\mathcal{T}} = b_{\mathcal{C}} + w_{\mathcal{T}}\\ &\boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{S}} = \begin{bmatrix} t_{\mathcal{M}} & b_{\mathcal{M}} \\ b_{\mathcal{M}} & t_{\mathcal{M}}\end{bmatrix}\,, \quad \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{D}} = \begin{bmatrix} t_{\mathcal{M}} & 0 \\ 0 & t_{\mathcal{M}}\end{bmatrix}\\ &\mathbf{A} = \boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathfrak{D}} - \boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathfrak{S}}\\ &\boldsymbol{\Sigma}_{\mathfrak{S}} = \begin{bmatrix} t_{\mathcal{E}} & b_{\mathcal{C}} \\ b_{\mathcal{C}} & t_{\mathcal{T}}\end{bmatrix}\,, \quad \boldsymbol{\Sigma}_{\mathfrak{D}} = \begin{bmatrix} t_{\mathcal{E}} & 0 \\ 0 & t_{\mathcal{T}}\end{bmatrix}\\ &\beta_{\mathfrak{h}} = -\frac{1}{2}\frac{\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h}})}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h}})}\,,\ \gamma^2_{\mathfrak{h}} = -\frac{1}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{h}})}\\ &\alpha^2_{\mathfrak{h}} = \gamma^2_{\mathfrak{h}} + \beta^2_{\mathfrak{h}}\,.\end{aligned} \tag{27}$$

We can observe that the model is over-parametrized: scaling $b_{\mathcal{M}}, w_{\mathcal{M}}, b_{\mathcal{C}}, w_{\mathcal{E}}, w_{\mathcal{T}}$ by some $\xi \neq 0$ results in the same solution for the VΓ parameters, thus we can arbitrarily fix any one of these parameters. In the following we set $w_{\mathcal{M}} = 1$. According to the PLDA LLR, the terms $\mu_{\mathfrak{D}}$ and $\mu_{\mathfrak{S}}$ should be equal, and should be tied to the remaining parameters. In this work we enrich the calibration model by treating the terms $\mu_{\mathfrak{D}}$ and $\mu_{\mathfrak{S}}$ as free parameters. The rationale derives from two observations. 1) PLDA-based models often include a bias term (e.g. Pairwise Support Vector Machines [24, 25, 26] or discriminative PLDA [27]). These terms result in score shifts that are optimal for the training criterion, but may not result in well-calibrated scores. We can capture such behavior through a shared location parameter. 2) Our derivations do not consider dataset shifts or the equivalent effects of linear terms that appear in PSVM or discriminative PLDA scoring functions. The resulting distributions would become more complex, and we are not aware of closed-form solutions even for isotropic models. We can, however, assume that the shift is sufficiently small, so that we can still approximate score distributions with VΓ densities. To capture the different location for the two score distributions, we introduce an additional free location parameter that represents the location differences between target and non-target distributions. In practice, both 1) and 2) can be accounted for through two independent location parameters $\mu_{\mathfrak{S}}$ and $\mu_{\mathfrak{D}}$. Our score models thus depends on the 7 free parameters $\lambda, \mu_{\mathfrak{D}}, \mu_{\mathfrak{S}}, b_{\mathcal{M}}, b_{\mathcal{C}}, w_{\mathcal{E}}, w_{\mathcal{T}}$. If enrollment and test data are affected only by i.i.d. nuisance, we can reduce the number of parameters to 6, by tying $w_{\mathcal{E}} = w_{\mathcal{T}} \triangleq w_{\mathcal{C}}$. Since the model parameters can be interpreted as "effective" variances, we refer to this last model as VΓ-Var. Given a set of calibration scores $(\mathcal{S}_{\mathfrak{S}}, \mathcal{S}_{\mathfrak{D}})$, the model can be trained by maximizing the weighted likelihood

$$\frac{\zeta}{|\mathcal{S}_{\mathfrak{S}}|}\sum_{s\in\mathcal{S}_{\mathfrak{S}}} f_{\mathcal{L}|\mathfrak{S}}(s) + \frac{1-\zeta}{|\mathcal{S}_{\mathfrak{D}}|}\sum_{s\in\mathcal{S}_{\mathfrak{D}}} f_{\mathcal{L}|\mathfrak{D}}(s)\,, \tag{28}$$

where $f_{\mathcal{L}|\mathfrak{S}}(s)$ and $f_{\mathcal{L}|\mathfrak{D}}(s)$ are the VΓ densities for the target and non-target distributions (26) and ζ is a weighting factor.

5. Utterance-dependent calibration

According to our model, given a trial $\mathbf{z} = \left[\boldsymbol{\phi}^T_{\mathcal{E}}\boldsymbol{\phi}^T_{\mathcal{T}}\right]^T$, the corresponding score is a sample of VΓ distributions in (26), where the parameters $w_{\mathcal{E}}$ and $w_{\mathcal{T}}$ represent "effective" within-class variability for the enrollment and test speaker vectors. To account for utterance-dependent nuisance, we can then assume that the terms $w_{\mathcal{E}}$ and $w_{\mathcal{T}}$ are not fixed, but vary from utterance

Table 1: *Calibration results on SRE 2019*

	PLDA			NL-PLDA			PSVM		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs[†]	0.192	0.418	5.2%	0.183	0.391	5.0%	0.155	0.342	4.0%
Log-Reg [1, 2]	0.200	0.438	5.2%	0.188	0.409	5.0%	0.165	0.360	4.0%
Linear VΓ [14]	0.206	0.424	5.2%	0.194	0.395	5.0%	0.169	0.347	4.0%
VΓ-Var	0.193	0.418	5.2%	0.184	0.392	5.0%	0.157	0.362	4.0%
Log-Reg + QM_4 [8]	0.190	0.434	4.8%	0.176	0.409	4.6%	0.153	0.374	3.7%
VΓ-Var + Dur	0.183	0.419	4.8%	0.174	0.388	4.6%	0.145	0.357	3.7%

[†] Minimum C_{llr}, minimum C_{prim}, and EER computed on the classifiers output

to utterance. In particular, for each speaker-vector $\boldsymbol{\phi}_i$ the effective within-class variance is w_i, and is a function of both i.i.d. and utterance-dependent miscalibration sources.

Given the significant effect of utterance duration variability on the accuracy of speaker verification systems, in this section we focus on modeling the effects of utterance duration on w_i, taking inspiration from i-vector [20] models. The i-vector model allows accounting for i-vector uncertainty through the i-vector posterior covariance matrix [28, 29, 30]. Incorporating the i-vector uncertainty at trial level is equivalent to modeling target and non-target trials as in (7) and (8), but replacing the covariance matrices $\mathbf{T}_\mathcal{E}$ and $\mathbf{T}_\mathcal{T}$ with $\mathbf{T}_\mathcal{E} = \mathbf{B}_\mathcal{C} + \mathbf{W}_\mathcal{E} + \mathbf{C}_{\mathcal{E},i}$ and $\mathbf{T}_\mathcal{T} = \mathbf{B}_\mathcal{C} + \mathbf{W}_\mathcal{T} + \mathbf{C}_{\mathcal{T},i}$, where $\mathbf{C}_{\mathcal{E},i}$ and $\mathbf{C}_{\mathcal{T},i}$ are the i-vector posterior covariances for enroll and test i-vectors of trial $\mathbf{z}_i$. An i-vector posterior covariance matrix $\mathbf{C}$ has a complex expression that depends on the zero-order statistics for an utterance computed on a Universal Background Model (UBM). However, it can be reasonably approximated [31] by a matrix whose form is

$$\mathbf{C} \approx (\mathbf{I} + D\mathbf{M})^{-1} , \tag{29}$$

where $\mathbf{M}$ depends on the UBM and the i-vector model parameters, whereas D is the utterance duration in frames. We further assume that $\mathbf{\Lambda}$ and $\mathbf{C}$ have the same principal directions, so that they can be both jointly diagonalized (a similar approximation was used in [32]). Each component $\mathbf{C}_j$ of $\mathbf{C}$ has then a functional form $\mathbf{C}_j = \frac{1}{1+D\eta_j} = \frac{\eta_j}{\eta_j+D}$. In this sense, we can interpret the i-vector posterior covariance matrix as a measure of the effects of utterance duration on the within-class variability of speaker vectors. To incorporate utterance duration in our score model, we adopt a similar functional relationship. We represent the effective variances $w_{\mathcal{E},i}$ and $w_{\mathcal{T},i}$ of a trial $\mathbf{z}_i$ as

$$w_{\mathcal{E},i} = w_\mathcal{E} + \frac{\psi}{D_{\mathcal{E},i} + \eta} , \quad w_{\mathcal{E},i} = w_\mathcal{T} + \frac{\psi}{D_{\mathcal{T},i} + \eta} , \tag{30}$$

where $D_{\mathcal{E},i}$ and $D_{\mathcal{T},i}$ are the enroll and test segment duration, and ψ and η are additional free model parameters, shared for all trials, that can be estimated by ML. As for the VΓ-Var model, also in this case we can assume[3] that $w_\mathcal{E} = w_\mathcal{T} = w_\mathcal{C}$. We refer to this model as VΓ-Var + Dur. It is worth noting that this approach can also be interpreted as a way to model speaker vector uncertainty in those cases where we have no access to uncertainty estimates (e.g. x-vectors), or uncertainty cannot be taken into account at classification level (e.g. PSVM).

[3] In some cases having different $w_\mathcal{E}$ and $w_\mathcal{T}$ can be useful to model differences in enrollment and test populations. Due to lack of space we do not further investigate the corresponding model in this work.

6. Experiments

We report results on the SRE 2019 [33] Evaluation set with three backends: PLDA with length normalization, Non-Linear PLDA (NL-PLDA) [34, 35] and Pairwise Support Vector Machines (PSVM) [24, 25]. Calibration parameters were estimated on a subset of the SRE 2019 Progress set. The front-end consists of a Deep Neural Network (DNN) with the same topology as in [36]. Details can be found in [14]. To assess the quality of our generative models, we consider a Logistic Regression baseline. For duration modeling, our baseline follows the approach of [8], which enriches the linear Log-Reg calibration model with quality measures (QM) that account for the effects of duration. In particular, we select QM_4 of [8], since it provided the best calibration results in our scenario. The results are reported in Table 1 in terms of C_{llr} [4, 37, 38] and actual primary cost C_{prim}. Since duration modeling improves discrimination, we also report Equal Error Rate (EER). The target prior for Log-Reg training and for the ML weight ζ was set to 0.1. The minor differences with respect to the results in [34] are due to slightly different backend training lists. Results show that the VΓ-Var model is effective, and provides close to optimal calibration for all classifiers. The results are similar to those of the NIG approach [13] we reported in [14]. Indeed, both models are able to capture non-linear miscalibration effects, and thus provide slightly better calibration, for supervised tasks, than linear models. The last two rows show that our approach is able to effectively account also for additional miscalibration sources such as utterance duration. Indeed, we can observe that VΓ-Var + Dur improves performance not only with respect to the other duration-agnostic models, but also with respect to minimum costs computed on raw scores. Compared to QM models, our approach achieves better performance with PLDA and PSVM, and similar C_{llr}, but better C_{prim}, with NL-PLDA.

7. Conclusions

We have presented a generative model able to incorporate utterance-dependent miscalibration sources in terms of "effective", non i.i.d., utterance-dependent within-class variance. This allows us, for example, to explicitly model utterance duration. The resulting generative model improves both calibration and verification accuracy, and achieves similar or better performance with respect to discriminative approaches based on quality measures. Being generative, the model can be extended to deal with missing labels. Future work will investigate the effectiveness for semi-supervised scenarios. Furthermore, our approach provides strong interpretations for the calibration parameters. We believe this to be an important step towards a unified model for score normalization and calibration.

8. References

[1] N. Brummer, L. Burget, and al., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 7, pp. 2072–2084, 2007. [Online]. Available: https://doi.org/10.1109/TASL.2007.902870

[2] N. Brümmer and G. R. Doddington, "Likelihood-ratio calibration using prior-weighted proper scoring rules," in *Proceedings of Interspeech*, 2013, pp. 1976–1979.

[3] N. Brümmer, "Focal toolkit," Available at http://sites.google.com/site/nikobrummer/focal.

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

[5] D. Ramos Castro, "Forensic evaluation of the evidence using automatic speaker recognition systems," Ph.D. dissertation, Autonomous University of Madrid, 2007.

[6] M. I. Mandasari, M. Gnther, R. Wallace, R. Saeidi, S. Marcel, and D. A. van Leeuwen, "Score calibration in face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 246–256, 2014.

[7] L. Ferrer, M. K. Nandwana, M. McLaren, D. Castan, and A. Lawson, "Toward fail-safe speaker recognition: Trial-based calibration with a reject option," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 140–153, 2019.

[8] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

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

[10] A. Nautsch, R. Saeidi, C. Rathgeb, and C. Busch, "Robustness of quality-based score calibration of speaker recognition systems with respect to low-snr and short-duration conditions," in *Proceedings of Odyssey 2016*, 2016.

[11] D. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proceedings of Interspeech*, 2013, pp. 1619–1623.

[12] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1680–1684.

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

[14] S. Cumani, "On the distribution of speaker verification scores: Generative models for unsupervised calibration," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 547–562, 2021.

[15] S. Cumani and P. Laface, "Tied normal variance–mean mixtures for linear score calibration," in *Proceedings of ICASSP 2019*, 05 2019, pp. 6121–6125.

[16] S. Cumani, "Normal variance-mean mixtures for unsupervised score calibration," in *Proceedings of Interspeech 2019*, 09 2019, pp. 401–405.

[17] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the 9th European Conference on Computer Vision*, ser. ECCV'06, vol. Part IV, 2006, pp. 531–542.

[18] P. Kenny, "Bayesian speaker verification with Heavy-Tailed Priors," in *Keynote presentation, Odyssey 2010, The Speaker and Language Recognition Workshop*, 2010.

[19] S. Cumani and P. Laface, "Generative pairwise models for speaker recognition," in *Proceedings of Odyssey 2014*, 2014.

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

[21] S. Cumani and P. Laface, "Speaker recognition using e–vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 736–748, 2018.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X–vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of ICASSP 2018*, 2018, pp. 5329–5333.

[23] D. Madan, P. Carr, and E. Chang, "The Variance Gamma process and option pricing," *Eurepean Finance Review*, vol. 2, pp. 79–105, 1998.

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

[25] S. Cumani and P. Laface, "Large scale training of Pairwise Support Vector Machines for speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1590–1600, 2014.

[26] S. Cumani, O. Glembek, N. Brümmer, E. de Villiers, and P. Laface, "Gender independent discriminative speaker recognition in i-vector space," in *Proceedings of ICASSP 2012*, 2012.

[27] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, and N. Brümmer, "Discriminatively trained Probabilistic Linear Discriminant Analysis for speaker verification," in *Proceedings of ICASSP 2011*, 2011, pp. 4832–4835.

[28] 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*, vol. 22, no. 4, pp. 846–857, 2014.

[29] S. Cumani, O. Plchot, and P. Laface, "Probabilistic Linear Discriminant Analysis of i–vector posterior distributions," in *Proceedings of ICASSP 2013*, 2013, pp. 7644–7648.

[30] P. Kenny, T. Stafylakis, P. Ouellet, M. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proceedings of ICASSP 2013*, 2013, pp. 7649–7653.

[31] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proceedings of ICASSP 2011*, 2011, pp. 4516–4519.

[32] S. Cumani, "Fast scoring of full posterior PLDA models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2036–2045, 2015.

[33] "The NIST 2019 speaker recognition evaluation: Cts challenge," 2012, available at https://www.nist.gov/system/files/documents/2019/07/22/2019_nist_speaker_recognition_challenge_v8.pdf.

[34] S. Cumani and P. Laface, "Joint estimation of PLDA and non–linear transformations of speaker vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1890–1900, 2017.

[35] S. Cumani and P. Laface, "Nonlinear i–vector transformations for PLDA–based speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 908–919, 2017.

[36] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5796–5800.

[37] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch University, South Africa, 2010.

[38] D. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," *Lecture Notes in Computer Science*, vol. 4343, pp. 330–353, 01 2007.