The Distributions of Uncalibrated Speaker Verification Scores: A Generative Model for Domain Mismatch and Trial-Dependent Calibration

The distributions of uncalibrated speaker verification scores: a generative model for domain mismatch and trial-dependent calibration

Sandro Cumani, Salvatore Sarni

***Abstract*—Speaker verification systems that compute log-likelihood ratios (LLR) between the same and different speaker hypotheses allow for cost-effective decisions that depend only on prior information. Domain mismatch, inaccurate model assumptions or the intrinsic nature of non-probabilistic classifiers often result in mis-calibrated scores, and a re-calibration step is required to map the classifier outputs to well-calibrated LLRs. Standard calibration is based on Logistic Regression, often paired with quality measures to provide trial-dependent calibration transformations. More recently, generative methods have been proposed as an alternative to discriminative approaches, which, however, are not yet able to exploit additional side information. In this work we introduce a novel generative approach based on the analysis of the effects of speaker vector distribution mismatch on the distribution of verification scores for PLDA and PLDA-based classifiers. We show that target and non-target scores can be modeled by Variance-Gamma distributions, whose parameters represent effective between and within-class variability. This allows us to introduce utterance-dependent variability models that can incorporate both explicit quality measures, such as the utterance duration, or implicit measures, such as the norm of a speaker embedding. Experimental results on different test sets with different front-ends and classifiers show that the proposed approach improves both calibration and verification accuracy with respect to state-of-the-art calibration models.**

***Index Terms*—Speaker verification, score calibration, log-likelihood ratio, duration variability, variance-gamma distribution**

I. Introduction

Speaker verification systems output scores that allow classifying sets of speech segments as belonging to the same speaker (target trial) or to different speakers (non-target trial). Cost-effective decisions require estimation of an application-dependent threshold, which, for well-calibrated scores, depends only on prior probabilities and on the costs of different kinds of errors. In many cases, however, the intrinsic nature of the classifier, mismatch between the training and evaluation populations or imprecise model assumptions may lead to poor calibration, which results in lowered performance of the recognizer. Score calibration approaches are usually employed to transform the recognizer outputs to approximately well-calibrated LLRs. The standard calibration approach is based on discriminative prior-weighted Logistic Regression (Log-Reg) [1], [2], successfully employed in a plethora of different scenarios [3]–[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]–[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 (CMLG). The model was further extended in [12] to handle missing labels. In [13] the authors proposed to model target and non-target scores with different, unconstrained densities, including T–student and Normal Inverse Gaussian (NIG). More recently, we have proposed a generative linear model [14]–[16] based on an analysis of the theoretical distribution of well-calibrated log-likelihood ratios of a Probabilistic Linear Discriminant Analysis (PLDA) [17], [18] classifier. We have shown that Variance-Gamma ($\mathrm{V}\Gamma$) distributions are suited for modeling well-calibrated scores, and we have provided sufficient conditions for Variance-Gamma distributed random variables to represent well-calibrated LLRs. We have also introduced the Constrained Maximum Likelihood Variance-Gamma (CMLVG) model, that assumes that target and non-target scores are obtained as an affine transformation of well-calibrated scores, modeled through $\mathrm{V}\Gamma$-distributed random variables, with parameters tied to satisfy the LLR constraint [11], [14]. The simple calibration model, paired with an accurate characterization of the score distribution, also allows for effective unsupervised training. However, as for standard logistic regression, it may provide sub-optimal results for those scenarios where linear functions do not provide a good approximation of optimal calibration transformations.

Both standard Log-Reg and generative models provide global calibration: scores are mapped to LLRs through a transformation that depends only on the score. In many scenarios, however, different trials may be affected by different miscalibration sources, and may thus require trial-dependent transformations. A well-known example is utterance duration: global calibration models tailored to long segments usually are not effective for short utterances. Discriminative approaches have been extended to partially handle trial-dependent miscalibration, for example by incorporating in the calibration transformation additional side-information, such as utterance duration or noise level [8]–[10], [19]. On the other hand, current generative models are not able to explicitly account for this kind of information. In this work we extend our analysis

The authors are with the Dipartimento di Automatica e Informatica, Politecnico di Torino, 10129 Torino, Italy (e-mail: sandro.cumani@polito.it, salvatore.sarni@polito.it).

of the distribution of well-calibrated PLDA scores [14] to explicitly account for distribution mismatches between training and evaluation populations. This allows us to derive a non-linear, generative score model that is able to characterize the distribution of target and non-target score in terms of VΓ densities whose parameters represent effective between and within-class variability of the training and evaluation speaker vectors (for example neural embeddings [20]). We then extend the model to incorporate trial-level mismatch sources. We present a simple duration model that can be combined with the score model to achieve state-of-the-art calibration for utterances of variable duration. Furthermore, we show that speaker vector-derived information, such as the squared norm of an utterance embedding, can be employed as a measure of utterance-level uncertainty, and incorporated at calibration level. Preliminary results with duration models were reported in [21]. In this work, we extend our analysis, providing a more detailed description of the theoretical framework and investigating additional side-information beyond utterance duration. We also significantly extend our experimental analysis.

II. Calibration models

Well-calibrated speaker verification systems compute the log-likelihood ratio for the evidence e associated to a trial:

$$x = LLR(e) = \log \frac{P(e|\mathfrak{S}, \mathcal{M})}{P(e|\mathfrak{D}, \mathcal{M})} \ , \tag{1}$$

where e represents a trial (e.g., a pair of i-vectors [22] or speaker embeddings [20]), $\mathcal{M}$ is a statistical model for e, and $\mathfrak{S}$ and $\mathfrak{D}$ are the target and non-target trial hypotheses, respectively. Usually, the statistical model $\mathcal{M}$ is estimated over a training population, and may not properly characterize the evaluation population in presence of domain mismatch. In other cases, a recognizer may be intrinsically unable to provide a well-calibrated LLR. For example, the Pairwise Support Vector Machine (PSVM) approach [23], [24] produces accurate scores which do not have a probabilistic interpretation. Score calibration is employed to estimate a mapping f_{cal} from an uncalibrated score s to a well-calibrated score $x = f_{cal}(s)$.

The standard discriminative approach for score calibration is prior-weighted logistic regression (Log-Reg) [1], [2]. Log-Reg optimizes the logarithmic proper scoring rule for a specific target prior assuming an affine calibration model, and has been extended to incorporate side information [8]–[10] through additional terms that depend on quality measures, such as the utterance duration:

$$f_{cal}^{QM}(s) = as + b + Q(q_{\mathcal{E}}, q_{\mathcal{T}}, \mathbf{q}) \tag{2}$$

where $q_{\mathcal{E}}, q_{\mathcal{T}}$ are the enrollment and test quality measures, and a, b and $\mathbf{q}$ are the model parameters. This approach allows modeling, for example, effects due to different utterance duration or noise levels. We will refer to these models as Logistic Regression with Quality Measures (LogReg + QM).

Alternative to discriminative methods, generative models estimate the calibration transformation f_{cal} through a statistical model $\mathcal{M}'$ that describes the distribution of the observed scores. These models interpret an observed score s as a sample of a Random Variable (R.V.) S, whose conditional densities given target and non-target hypotheses are $f_{S|\mathfrak{S}}$ and $f_{S|\mathfrak{D}}$, respectively[1]. Given a score s, the calibration function corresponds to the LLR for the score under the two hypotheses:

$$x' = f_{cal}(s) = \log \frac{f_{S|\mathfrak{S}}(s)}{f_{S|\mathfrak{D}}(s)} \ . \tag{3}$$

The statistical model can either consist of an explicit expression for the target and non-target densities, which implicitly defines the calibration transformation, or be provided through an explicit model for the calibration transformation f_{cal} and a model for the distribution of the calibrated scores. In the latter case, the distributions of the calibrated scores are constrained as to satisfy the LLR property [11]. For example, constrained Gaussian densities paired with a linear calibration model (CMLG) have been proposed in [11]. In [14] we have shown that Variance-Gamma (VΓ) distributions provide more accurate characterization of the scores generated by PLDA and PLDA-derived models. Well-calibrated LLRs were thus assumed to be samples of VΓ-distributed R.V.s,

$$X|\mathfrak{D} \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta, \mu) \ , \quad X|\mathfrak{S} \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta + 1, \mu) \ , \tag{4}$$

where the VΓ density is defined as

$$f_{\mathrm{V\Gamma}}(x|\lambda, \alpha, \beta, \mu) = \frac{\gamma^{2\lambda}|x-\mu|^{\lambda-\frac{1}{2}} K_{\lambda-\frac{1}{2}}(\alpha|x-\mu|)}{\sqrt{\pi}\Gamma(\lambda)(2\alpha)^{\lambda-\frac{1}{2}}} e^{\beta(x-\mu)} \tag{5}$$

with $\lambda > 0, \alpha > |\beta|$ and $\gamma^2 = \alpha^2 - \beta^2$, and $K_\nu(x)$ denotes the modified Bessel function of the third kind of order ν. The parameters of the distributions in (4), $\mu, \beta, \alpha > \max(|\beta|, |\beta+1|), \lambda > 0$, are tied between the target and non-target densities as to satisfy the LLR constraint. As for CMLG, the model can be paired with an affine calibration transformation to obtain the distribution of the observed scores. The parameters of the calibration transformation and of the well-calibrated distributions can be estimated by Maximum Likelihood. The resulting model is effective for tasks where linear calibration is sufficient. Furthermore, the more accurate characterization of the scores, combined with a simple calibration transformation, allows also for robust unsupervised estimation of the model parameters. We will refer to this model as Linear VΓ.

While effective at compensating global miscalibration, generative approaches are not able to account for trial-dependent miscalibration sources, such as utterance duration. Furthermore, both discriminative and generative models rely on assumptions that are not directly related to the distributions of speaker vectors. In the following sections we therefore analyze the effects of domain mismatch on the distribution of scores of PLDA classifiers.

III. The distribution of PLDA scores

The simplified PLDA[2] model [26] assumes that a speaker vector is a realization of a Random Vector (R.V.) $\mathbf{\Phi}$ that can be expressed as the sum of two factors

$$\mathbf{\Phi} = \overline{\mathbf{Y}} + \overline{\mathbf{E}} \ . \tag{6}$$

[1]For the sake of readability we omit conditioning $f_{S|\mathfrak{S}}$ and $f_{S|\mathfrak{D}}$ on the model $\mathcal{M}'$

[2]Extension to subspace models is straightforward — see, for example, [25]

$\mathbf{\Phi}$ is the M-dimensional R.V. responsible for generating an observed speaker vector $\boldsymbol{\phi}$ (e.g. an i-vector [22], e-vector [27], or speaker embedding [20]), $\overline{\mathbf{Y}}$ is the R.V. representing the speaker identity and $\overline{\mathbf{E}}$ represents residual noise. The model assumes that $\overline{\mathbf{Y}}$ and $\overline{\mathbf{E}}$ are a priori normal distributed, as:

$$\overline{\mathbf{Y}} \sim \mathcal{N}(\mathbf{m}_{\mathcal{M}}, \mathbf{B}_{\mathcal{M}}) \ , \quad \overline{\mathbf{E}} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_{\mathcal{M}}) \ . \tag{7}$$

$\mathbf{m}_{\mathcal{M}}$ represents the global dataset mean, whereas $\mathbf{B}_{\mathcal{M}}$ and $\mathbf{W}_{\mathcal{M}}$ can be interpreted as between-class and within class covariance matrices. Without loss of generality[3]we also assume $\mathbf{m}_{\mathcal{M}} = 0$, and that both $\mathbf{B}_{\mathcal{M}}$ and $\mathbf{W}_{\mathcal{M}}$ are diagonal. The PLDA score for a pair of speaker vectors $\mathbf{z} = \left[\boldsymbol{\phi}_{\mathcal{E}}^T, \boldsymbol{\phi}_{\mathcal{T}}^T\right]^T$ is

$$\ell(\mathbf{z}) = K_{\mathcal{M}} - \frac{1}{2}\mathbf{z}^T \left(\mathbf{\Sigma}_{\mathcal{M},\mathfrak{S}}^{-1} - \mathbf{\Sigma}_{\mathcal{M},\mathfrak{D}}^{-1}\right)\mathbf{z} \ , \tag{8}$$

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

where $\mathbf{T}_{\mathcal{M}} = \mathbf{B}_{\mathcal{M}} + \mathbf{W}_{\mathcal{M}}$ and $K_{\mathcal{M}}$ collects the constant terms. Since $\mathbf{B}_{\mathcal{M}}, \mathbf{W}_{\mathcal{M}}$ are diagonal, it can be expressed as a sum of M terms

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

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

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

$$\mathbf{T}_{\mathcal{M}_i} = \mathbf{B}_{\mathcal{M},i} + \mathbf{W}_{\mathcal{M},i} \ , \tag{11}$$

$$k_i = \frac{1}{2}\log \mathbf{T}_{\mathcal{M},i}^2 - \frac{1}{2}\log\left(\mathbf{T}_{\mathcal{M},i}^2 - \mathbf{B}_{\mathcal{M},i}^2\right) \ . \tag{12}$$

PLDA assumes that (6) holds for both the training population, that is used to estimate the values of $\mathbf{B}_{\mathcal{M}}$ and $\mathbf{W}_{\mathcal{M}}$, and for the evaluation population. However, in presence of domain mismatch the evaluation population may be better characterized by a different model. The distribution of the evaluation population affects the distribution of the PLDA scores and, in general, mismatches result in non-calibrated scores. In the following we assume that evaluation trials are realizations of a PLDA model that differs from the model used to compute the score. In particular, we assume that an evaluation trial is a sample of R.V. $\mathbf{Z}$, with conditional distributions given by

$$\begin{bmatrix} \mathbf{\Phi}_{\mathcal{E}} \\ \mathbf{\Phi}_{\mathcal{T}} \end{bmatrix} | \mathfrak{D} \sim \mathcal{N}\left(\mathbf{m}, \mathbf{\Sigma}_{\mathfrak{D}}\right) \ , \mathbf{\Sigma}_{\mathfrak{D}} = \begin{bmatrix} \mathbf{T}_{\mathcal{E}} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{\mathcal{T}} \end{bmatrix} \ ,$$
$$\begin{bmatrix} \mathbf{\Phi}_{\mathcal{E}} \\ \mathbf{\Phi}_{\mathcal{T}} \end{bmatrix} | \mathfrak{S} \sim \mathcal{N}\left(\mathbf{m}, \mathbf{\Sigma}_{\mathfrak{S}}\right) \ , \mathbf{\Sigma}_{\mathfrak{S}} = \begin{bmatrix} \mathbf{T}_{\mathcal{E}} & \mathbf{B}_{\mathcal{C}} \\ \mathbf{B}_{\mathcal{C}} & \mathbf{T}_{\mathcal{T}} \end{bmatrix} \ , \tag{13}$$

and $\mathbf{T}_{\mathcal{E}} = \mathbf{B}_{\mathcal{C}} + \mathbf{W}_{\mathcal{E}}$ and $\mathbf{T}_{\mathcal{T}} = \mathbf{B}_{\mathcal{C}} + \mathbf{W}_{\mathcal{T}}$, corresponding to a PLDA model

$$\mathbf{\Phi}_{\mathcal{S}} = \mathbf{Y}_{\mathcal{C}} + \mathbf{E}_{\mathcal{S}} \ , \quad \mathcal{S} \in \{\mathcal{E}, \mathcal{T}\} \ , \tag{14}$$
$$\mathbf{Y}_{\mathcal{C}} \sim \mathcal{N}(\mathbf{m}, \mathbf{B}_{\mathcal{C}}) \ , \quad \mathbf{E}_{\mathcal{S}} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_{\mathcal{S}})$$

[3]These assumptions can be recovered through an affine transformation of the data.

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, which may also differ across different segments (e.g. due to the intrinsic difference in the amount of information contained in a spoken segment, or to the uncertainty in the estimation of different speaker vectors [28]–[30]). To keep the model tractable, we assume that $\mathbf{B}_{\mathcal{C}}$, $\mathbf{W}_{\mathcal{E}}$ and $\mathbf{W}_{\mathcal{T}}$ are diagonal. Although this may seem a strong assumption, experimental results confirm that the resulting model is powerful enough to improve calibration with respect to state-of-the-art models. We also assume[4] $\mathbf{m} = \mathbf{0}$. As we showed in [14], if an evaluation trial $\mathbf{z}$ is a realization of R.V. $\mathbf{Z}$, its score can be interpreted as a realization of R.V. $\mathcal{L} = \ell(\mathbf{Z})$. To derive the distribution of target and non-target scores, we consider R.V.s[5]

$$\mathbf{Z}_{\mathfrak{D}} \sim \mathbf{Z}|\mathfrak{D} \ , \quad \mathbf{Z}_{\mathfrak{S}} \sim \mathbf{Z}|\mathfrak{S} \ .$$

The scores are therefore realizations of two R.V.s

$$\mathcal{L}|\mathfrak{D} = \ell(\mathbf{Z}|\mathfrak{D}) \ , \quad \mathcal{L}|\mathfrak{S} = \ell(\mathbf{Z}|\mathfrak{S}) \tag{15}$$

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 different 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, with

$$\mathbf{Z}_i|\mathfrak{D} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\mathfrak{D},i}\right) \ , \quad \mathbf{Z}_i|\mathfrak{S} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\mathfrak{S},i}\right) \ , \tag{16}$$

where

$$\mathbf{\Sigma}_{\mathfrak{D},i} = \begin{bmatrix} \mathbf{T}_{\mathcal{E},i} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{\mathcal{T},i} \end{bmatrix} \ , \quad \mathbf{\Sigma}_{\mathfrak{S},i} = \begin{bmatrix} \mathbf{T}_{\mathcal{E},i} & \mathbf{B}_{\mathcal{C},i} \\ \mathbf{B}_{\mathcal{C},i} & \mathbf{T}_{\mathcal{T},i} \end{bmatrix} \ . \tag{17}$$

$\mathbf{B}_{\mathcal{C},i}$, $\mathbf{T}_{\mathcal{E},i}$ 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. From (9), the R.V.s that describe the distribution of target and non-target PLDA scores can be expressed as sums of M independent R.V.s as:

$$\mathcal{L}|\mathfrak{h} = \ell(\mathbf{Z}|\mathfrak{h}) = \sum_{i=1}^{M} \mathcal{L}_i|\mathfrak{h} \ , \quad \mathfrak{h} \in \{\mathfrak{S}, \mathfrak{D}\} \ , \tag{18}$$

$$\mathcal{L}_i|\mathfrak{h} = \ell(\mathbf{Z}_i|\mathfrak{h}) = \frac{1}{2}\mathbf{Z}_{\mathfrak{h},i}^T \mathbf{A}_i \mathbf{Z}_{\mathfrak{h},i} + k_i \ , \tag{19}$$

$$\mathbf{Z}_{\mathfrak{h},i} \sim \mathbf{Z}_i|\mathfrak{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathfrak{h},i}) \ . \tag{20}$$

Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a standard normal distributed R.V. Let also $\mathbf{C}_{\mathfrak{h},i}$ be the Cholesky decomposition of $\mathbf{\Sigma}_{\mathfrak{h},i} = \mathbf{C}_{\mathfrak{h},i}\mathbf{C}_{\mathfrak{h},i}^T$, and $\mathbf{U}_{\mathfrak{h},i}\mathbf{D}_{\mathfrak{h},i}\mathbf{U}_{\mathfrak{h},i}^T$ denote the eigen-decomposition of

$$\mathbf{M}_{\mathfrak{h},i} \triangleq \mathbf{C}_{\mathfrak{h},i}^T \mathbf{A}_i \mathbf{C}_{\mathfrak{h},i} = \mathbf{U}_{\mathfrak{h},i}\mathbf{D}_{\mathfrak{h},i}\mathbf{U}_{\mathfrak{h},i}^T \ . \tag{21}$$

[4]The evaluation population mean can be easily estimated and compensated from few, unlabeled evaluation samples.

[5]If $X(\omega)$ is a R.V. defined over the probability space $(\Omega, \mathcal{A}, \mathcal{P})$, and H is an event, then $X|H$ is not a R.V. over $(\Omega, \mathcal{A}, \mathcal{P}(\cdot))$. However, we can interpret conditioning as acting on the probability space rather than just on the probability distribution [31]. In this sense, we can define $(X|H)(\omega)$ as a R.V. which is functionally identical to $X(\omega)$, $(X|H)(\omega) := X(\omega)$, but is defined on the probability space $(\Omega \cap H, \{H \cap A : A \in \mathcal{A}\}, \mathcal{P}(\cdot,H)/\mathcal{P}(H))$. As long as we don't mix R.V.s defined over different spaces, we can work with the R.V.s $\mathbf{Z}_{\mathfrak{D}} \sim \mathbf{Z}|\mathfrak{D}$ and $\mathbf{Z}_{\mathfrak{S}} \sim \mathbf{Z}|\mathfrak{S}$, each defined on its own probability space, whose distributions correspond to the conditional distributions of R.V. $\mathbf{Z}$, given the labeling hypotheses $\mathfrak{h}$. With an abuse of notation, we will denote both the conditional distribution of $\mathbf{Z}|\mathfrak{h}, \mathfrak{h} \in \{\mathfrak{D}, \mathfrak{S}\}$ and the distribution of R.V. $\mathbf{Z}_{\mathfrak{h}}$ with the same symbol $\mathbf{Z}|\mathfrak{h}$. The same considerations apply to score distributions $\mathcal{L}|\mathfrak{h}$.

We can express the conditional distributions of $\mathcal{L}_i$ as

$$\mathcal{L}_i|\mathfrak{h} \sim \frac{1}{2}\mathbf{X}^T\mathbf{C}_{\mathfrak{h},i}^T\mathbf{A}_i\mathbf{C}_{\mathfrak{h},i}\mathbf{X}+k_i \sim \frac{1}{2}\mathbf{Y}_{\mathfrak{h}}^T\mathbf{D}_{\mathfrak{h},i}\mathbf{Y}_{\mathfrak{h}}+k_i \ , \quad (22)$$

where

$$\mathbf{Y}_{\mathfrak{h}} = \mathbf{U}_{\mathfrak{h},i}\mathbf{X} \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \quad (23)$$

Since the determinant of $\mathbf{A}_i$ is negative, the determinant of $\mathbf{M}_{\mathfrak{h},i}$ is also negative, thus its two eigenvalues have different sign. To derive an expression for $\mathcal{L}_i$ we analyze the distribution of quadratic, indefinite forms (22). For the sake of readability, we drop all suffices and we consider quadratic forms

$$\mathcal{L} = \frac{1}{2}\mathbf{Y}^T\mathbf{D}\mathbf{Y}+k = \frac{1}{2}d_+Y_+^2 - \frac{1}{2}|d_-|Y_-^2+k \quad (24)$$

where $\mathbf{D}$ is a 2×2 diagonal matrix with diagonal elements $d_+ > 0$ and $d_- < 0$ and $\mathbf{Y}$ is standard normal distributed:

$$\mathbf{D} = \begin{bmatrix} d_+ & 0 \\ 0 & d_- \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} Y_+ \\ Y_- \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right). \quad (25)$$

We observe that $Y_+$and Y_- are independent, and Y_+^2 and Y_-^2 are Gamma-distributed as

$$Y_+ \sim \Gamma\left(\frac{1}{2},\frac{1}{2}\right) \ , \quad Y_- \sim \Gamma\left(\frac{1}{2},\frac{1}{2}\right) \ . \quad (26)$$

$\mathcal{L}$ corresponds to a difference of independent, Gamma-distributed R.V.s

$$\mathcal{L} = G_+ - G_- + k \ , \quad G_+ \triangleq \frac{1}{2}d_+Y_+^2 \ , \quad G_- \triangleq \frac{1}{2}|d_-|Y_-^2 \ ,$$
$$G_+ \sim \Gamma\left(\frac{1}{2},\frac{1}{d_+}\right) \ , \quad G_- \sim \Gamma\left(\frac{1}{2},\frac{1}{|d_-|}\right) \ . \quad (27)$$

The Moment Generating Function (MGF) of $\mathcal{L}$ is

$$\begin{aligned} M_{\mathcal{L}}(t) &= e^{kt}M_{G_+}(t)M_{G_-}(-t) \\ &= e^{kt}(1-d_+t)^{-\frac{1}{2}}(1-d_-t)^{-\frac{1}{2}} \\ &= e^{kt}\big(1-\mathrm{Tr}(\mathbf{D})t+\det(\mathbf{D})t^2\big)^{-\frac{1}{2}} \ , \end{aligned} \quad (28)$$

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

$$M_X(t) = e^{\mu t}\left(1-\frac{2\beta}{\gamma^2}t-\frac{t^2}{\gamma^2}\right)^{-\lambda} \ , \quad (29)$$

where $\gamma^2 = \alpha^2-\beta^2$. Therefore, $\mathcal{L}$ is VΓ-distributed [14], [32], and the parameters can be recovered by inspection:

$$\lambda = \frac{1}{2} \ , \ \mu = k \ , \ \gamma^2 = -\frac{1}{\det(\mathbf{D})} \ , \ \beta = -\frac{1}{2}\frac{\mathrm{Tr}(\mathbf{D})}{\det(\mathbf{D})} \ . \quad (30)$$

Combining (30) and (22) we have that $\mathcal{L}_i|\mathfrak{D}$ and $\mathcal{L}_i|\mathfrak{S}$ are also VΓ-distributed. From (21), we also have that

$$\begin{aligned} \mathrm{Tr}(\mathbf{D}_{\mathfrak{h},i}) &= \mathrm{Tr}(\mathbf{M}_{\mathfrak{h},i}) = \mathrm{Tr}(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{h},i}) \\ \det(\mathbf{D}_{\mathfrak{h},i}) &= \det(\mathbf{M}_{\mathfrak{h},i}) = \det(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{h},i}) \end{aligned} \quad (31)$$

thus

$$\mathcal{L}_i|\mathfrak{h} \sim \mathrm{V}\Gamma\left(\frac{1}{2},\alpha_{\mathfrak{h},i},\beta_{\mathfrak{h},i},k_i\right) \quad (32)$$

with parameters

$$\begin{aligned} \beta_{\mathfrak{D},i} &= -\frac{1}{2}\frac{\mathrm{Tr}(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{D},i})}{\det(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{D},i})} \ , & \beta_{\mathfrak{S},i} &= -\frac{1}{2}\frac{\mathrm{Tr}(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{S},i})}{\det(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{S},i})} \ , \\ \gamma^2_{\mathfrak{D},i} &= -\frac{1}{\det(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{D},i})} \ , & \gamma^2_{\mathfrak{D},i} &= -\frac{1}{\det(\mathbf{A}_i\mathbf{\Sigma}_{\mathfrak{S},i})} \ , \\ \alpha^2_{\mathfrak{D},i} &= \gamma^2_{\mathfrak{D},i}+\beta^2_{\mathfrak{D},i} \ , & \alpha^2_{\mathfrak{S},i} &= \gamma^2_{\mathfrak{S},i}+\beta^2_{\mathfrak{S},i} \ . \end{aligned} \quad (33)$$

The parameters in (33) depend on the parameters of the speaker vectors distributions $\mathbf{B}_{\mathcal{M},i},\mathbf{W}_{\mathcal{M},i},\mathbf{B}_{\mathcal{C},i},\mathbf{W}_{\mathcal{E},i},\mathbf{W}_{\mathcal{T},i}$ only through the ratios

$$\eta_{\mathcal{E},i} = \frac{\mathbf{T}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{E},i}} \ , \quad \eta_{\mathcal{T},i} = \frac{\mathbf{T}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{T},i}} \ , \quad (34)$$

$$\rho_{\mathcal{E},i} = \frac{\mathbf{B}_{\mathcal{C},i}}{\mathbf{W}_{\mathcal{E},i}} \ , \quad \rho_{\mathcal{T},i} = \frac{\mathbf{B}_{\mathcal{C},i}}{\mathbf{W}_{\mathcal{T},i}} \ , \quad \rho_{\mathcal{M},i} = \frac{\mathbf{B}_{\mathcal{M},i}}{\mathbf{W}_{\mathcal{M},i}} \quad (35)$$

$\eta_{\mathcal{E},i}$ and $\eta_{\mathcal{T},i}$ represent the ratio between the scales of the training population and the enrollment and test population, respectively, whereas $\rho_{\mathcal{M},i},\rho_{\mathcal{E},i}$ and $\rho_{\mathcal{T},i}$ represent the between-to-within class variability ratios for the training, enrollment and test populations, respectively. We can observe that the parametrization is redundant, since $\eta_{\mathcal{E},i}$ and $\eta_{\mathcal{T},i}$ can be expressed in terms $\rho_{\mathcal{E},i}$ and $\rho_{\mathcal{T},i}$ and of a common ratio $\eta_{\mathcal{C},i} = \frac{2\mathbf{T}_{\mathcal{M},i}}{\mathbf{T}_{\mathcal{E},i}+\mathbf{T}_{\mathcal{T},i}}$ representing the ratio between the scale of the training population and the average scale of enrollment and test population. In many cases it's reasonable to assume that the enrollment and test populations are homogeneous, i.e. that $\mathbf{W}_{\mathcal{E}} = \mathbf{W}_{\mathcal{T}} \triangleq \mathbf{W}_{\mathcal{C}}$. In this case $\rho_{\mathcal{E},i} = \rho_{\mathcal{T},i} \triangleq \rho_{\mathcal{C},i}$, $\eta_{\mathcal{E},i} = \eta_{\mathcal{T},i} = \eta_{\mathcal{C},i}$ and after some algebraic manipulations we obtain

$$\begin{aligned} \gamma^2_{\mathfrak{D},i} &= \eta^2_{\mathcal{C},i}\frac{1+2\rho_{\mathcal{M},i}}{\rho^2_{\mathcal{M},i}} \ , & \gamma^2_{\mathfrak{S},i} &= \gamma^2_{\mathfrak{D},i}\frac{(1+\rho_{\mathcal{C},i})^2}{1+2\rho_{\mathcal{C},i}} \ , \\ \beta_{\mathfrak{D},i} &= -\eta_{\mathcal{C},i} \ , & \beta_{\mathfrak{S},i} &= \eta_{\mathcal{C},i}\frac{1+\rho_{\mathcal{C},i}}{1+2\rho_{\mathcal{C},i}}\left(\frac{\rho_{\mathcal{C},i}}{\rho_{\mathcal{M},i}}-1\right). \end{aligned} \quad (36)$$

We can observe that 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 shape of the non-target distribution does not depend on the between-over-within variability ratio $\rho_{\mathcal{C},i}$ of the evaluation population. The results for the distribution of well-calibrated PLDA scores we derived in [14] can be recovered by assuming that the evaluation population follows the same distribution as the training population, i.e. $\rho_{\mathcal{E},i} = \rho_{\mathcal{T},i} = \rho_{\mathcal{M},i}$ and $\eta_{\mathcal{E},i} = \eta_{\mathcal{T},i} = 1$.

As for the well-calibrated case, we are not aware of closed form solutions for the distribution of $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$. However, if we assume that the ratios $\eta_{\mathcal{C},i}$, $\rho_{\mathcal{M},i}$, $\rho_{\mathcal{E},i}$ and $\rho_{\mathcal{T},i}$ are the same for all directions i

$$\eta_{\mathcal{C},i} = \ \eta_{\mathcal{C}}, \ \ \rho_{\mathcal{M},i} = \rho_{\mathcal{M}}, \ \ \rho_{\mathcal{E},i} = \rho_{\mathcal{E}}, \ \ \rho_{\mathcal{T},i} = \rho_{\mathcal{T}}, \ \ \forall i \quad (37)$$

then $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ are again VΓ-distributed, with

$$\mathcal{L}|\mathfrak{h} \sim \mathrm{V}\Gamma\left(\frac{M}{2},\alpha_{\mathfrak{h}},\beta_{\mathfrak{h}},\sum_{i=1}^{M}k_i\right) \ , \quad (38)$$

where the parameters $\alpha_{\mathfrak{D}}, \beta_{\mathfrak{D}}, \alpha_{\mathfrak{S}}, \beta_{\mathfrak{S}}$ can be computed from (33) using the any index i.

IV. A GENERATIVE MODEL FOR MISMATCHED DATA

Equation (38) provides a model for the distribution of target and non-target scores, assuming that the M-dimensional speaker vectors are distributed according to (13), and that, for each speaker vector direction, the covariance matrices are related to a common set of normalized variances as

$$\begin{aligned} &\mathbf{B}_{\mathcal{M},i} = \xi_i b_{\mathcal{M}} \ , \mathbf{W}_{\mathcal{M},i} = \xi_i w_{\mathcal{M}} \\ &\mathbf{B}_{\mathcal{C},i} = \xi_i b_{\mathcal{C}} \ , \quad \mathbf{W}_{\mathcal{E},i} = \xi_i w_{\mathcal{E}} \ , \mathbf{W}_{\mathcal{T},i} = \xi_i w_{\mathcal{T}} \ , \end{aligned} \tag{39}$$

for scalars $\xi_i \neq 0$. Although assuming isotropic normalized variances may seem a strong hypothesis, we have shown in [14] that this approximation is effective as long as we assume that the score depends only on a small set of "effective" speaker vector dimensions. Indeed, most speaker vector dimensions, which have small between-over-within variability ratios, provide small contributions to the PLDA scores. We can therefore assume that the score has been generated by an effective speaker vector with smaller dimensionality, which approximately satisfies the assumption (39). In terms of the distributions of target and non-target scores, this can be modeled through the parameter λ of the VΓ distributions [14]. As starting point to derive our model we therefore consider the score model

$$\mathcal{L}|\mathfrak{D} \sim \mathrm{V\Gamma}\left(\lambda, \mu, \alpha_{\mathfrak{D}}, \beta_{\mathfrak{D}}\right) \ , \ \mathcal{L}|\mathfrak{S} \sim \mathrm{V\Gamma}\left(\lambda, \mu, \alpha_{\mathfrak{S}}, \beta_{\mathfrak{S}}\right) \tag{40}$$

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

$$\begin{aligned} &t_{\mathcal{M}} = b_{\mathcal{M}} + w_{\mathcal{M}} \ , \quad t_{\mathcal{E}} = b_{\mathcal{C}} + w_{\mathcal{E}} \ , \quad t_{\mathcal{T}} = b_{\mathcal{C}} + w_{\mathcal{T}} \\ &\mathbf{\Sigma}_{\mathcal{M},\mathfrak{S}} = \begin{bmatrix} t_{\mathcal{M}} & b_{\mathcal{M}} \\ b_{\mathcal{M}} & t_{\mathcal{M}} \end{bmatrix} \ , \qquad \mathbf{\Sigma}_{\mathcal{M},\mathfrak{D}} = \begin{bmatrix} t_{\mathcal{M}} & 0 \\ 0 & t_{\mathcal{M}} \end{bmatrix} \\ &\mathbf{A} = \mathbf{\Sigma}_{\mathcal{M},\mathfrak{D}}^{-1} - \mathbf{\Sigma}_{\mathcal{M},\mathfrak{S}}^{-1} \\ &\mathbf{\Sigma}_{\mathfrak{S}} = \begin{bmatrix} t_{\mathcal{E}} & b_{\mathcal{C}} \\ b_{\mathcal{C}} & t_{\mathcal{T}} \end{bmatrix} \ , \qquad \mathbf{\Sigma}_{\mathfrak{D}} = \begin{bmatrix} t_{\mathcal{E}} & 0 \\ 0 & t_{\mathcal{T}} \end{bmatrix} \\ &\beta_{\mathfrak{D}} = -\frac{1}{2}\frac{\mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}_{\mathfrak{D}})}{\det(\mathbf{A}\mathbf{\Sigma}_{\mathfrak{D}})} \ , \qquad \beta_{\mathfrak{S}} = -\frac{1}{2}\frac{\mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}_{\mathfrak{S}})}{\det(\mathbf{A}\mathbf{\Sigma}_{\mathfrak{S}})} \ , \\ &\gamma_{\mathfrak{D}}^2 = -\frac{1}{\det(\mathbf{A}\mathbf{\Sigma}_{\mathfrak{D}})} \ , \qquad \gamma_{\mathfrak{S}}^2 = -\frac{1}{\det(\mathbf{A}\mathbf{\Sigma}_{\mathfrak{S}})} \ , \\ &\alpha_{\mathfrak{D}}^2 = \gamma_{\mathfrak{D}}^2 + \beta_{\mathfrak{D}}^2 \ , \qquad \alpha_{\mathfrak{S}}^2 = \gamma_{\mathfrak{S}}^2 + \beta_{\mathfrak{S}}^2 \ . \end{aligned} \tag{41}$$

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 $\xi > 0$ results in the same solution for the VΓ parameters, thus we can arbitrarily fix any one of these parameters. In the following we assume $w_{\mathcal{M}} = 1$. According to the PLDA model, the location parameter μ is shared by both distributions, and should be tied to the remaining parameters. However, PLDA-derived models often include bias terms (e.g. Pairwise Support Vector Machines [23], [24], [33] or discriminative PLDA [34]). 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 that is independent from the remaining parameters and is estimated form the data. Furthermore, 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, and model its effects in terms of a linear transformation of the VΓ distributions. For this reasons, we introduce an additional free location parameter and an additional free scaling parameter that represent the location and scale differences between target and non-target distributions. In practice, both location and scaling parameters can be accounted for through two independent location parameters $\mu_{\mathfrak{S}}$ and $\mu_{\mathfrak{D}}$, and a scaling parameters $\mathbf{a}_{\mathfrak{S}}$ that scales the target distribution (we do not require an equivalent scaling parameter for the non-target class, since the model can already estimate the distribution scaling through the original parameters (39)).

The proposed VΓ model becomes

$$\mathcal{L}|\mathfrak{D} \sim \mathrm{V\Gamma}\left(\lambda, \mu_{\mathfrak{D}}, \alpha_{\mathfrak{D}}, \beta_{\mathfrak{D}}\right), \quad \mathcal{L}|\mathfrak{S} \sim \mathrm{V\Gamma}\left(\lambda, \mu_{\mathfrak{S}}, \frac{\alpha_{\mathfrak{S}}}{a_{\mathfrak{S}}}, \frac{\beta_{\mathfrak{S}}}{a_{\mathfrak{S}}}\right) \tag{42}$$

Our model depends on the 8 free parameters $\mathbf{\Pi} = (\lambda, \mu_{\mathfrak{D}}, \mu_{\mathfrak{S}}, b_{\mathcal{M}}, b_{\mathcal{C}}, w_{\mathcal{E}}, w_{\mathcal{T}}, a_{\mathfrak{S}})$. If enrollment and test data are affected only by i.i.d. nuisance, we can reduce the number of parameters to 7, by tying $w_{\mathcal{E}} = w_{\mathcal{T}} = w_{\mathcal{C}}$. Since the parameters can be interpreted as "effective" variance, we refer to this 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 [14]

$$\arg\max_{\mathbf{\Pi}} \frac{\zeta}{|\mathcal{S}_{\mathfrak{S}}|} \sum_{s \in \mathcal{S}_{\mathfrak{S}}} f_{\mathcal{L}|\mathfrak{S}}\left(s; \mathbf{\Pi}\right) + \frac{1-\zeta}{|\mathcal{S}_{\mathfrak{D}}|} \sum_{s \in \mathcal{S}_{\mathfrak{D}}} f_{\mathcal{L}|\mathfrak{D}}\left(s; \mathbf{\Pi}\right) \ , \tag{43}$$

where $f_{\mathcal{L}|\mathfrak{S}}\left(s; \mathbf{\Pi}\right)$ and $f_{\mathcal{L}|\mathfrak{D}}\left(s; \mathbf{\Pi}\right)$ are the VΓ densities for the target and non-target distributions (40) and ζ is a weighting factor. For a given set of parameters $\mathbf{\Pi} = (\lambda, \mu_{\mathfrak{D}}, \mu_{\mathfrak{S}}, b_{\mathcal{M}}, b_{\mathcal{C}}, w_{\mathcal{E}}, w_{\mathcal{T}}, a_{\mathfrak{S}})$ the calibration transformation is given by

$$f_{cal}(s; \mathbf{\Pi}) = \frac{\mathrm{V\Gamma}\left(s|\lambda, \mu_{\mathfrak{S}}, \frac{\alpha_{\mathfrak{S}}}{a_{\mathfrak{S}}}, \frac{\beta_{\mathfrak{S}}}{a_{\mathfrak{S}}}\right)}{\mathrm{V\Gamma}\left(s|\lambda, \mu_{\mathfrak{D}}, \alpha_{\mathfrak{D}}, \beta_{\mathfrak{D}}\right)} \tag{44}$$

with the VΓ densities defined in (5), and will, in general, consist of a non-linear function of the scores. Since the model defines a calibration transformation, we can also discriminatively estimate the model parameters through prior-weighted logistic regression. The discriminative objective function can be expressed as

$$\begin{aligned} \arg\min_{\mathbf{\Pi}} \quad & \frac{\pi}{|S_{\mathfrak{S}}|} \sum_{s \in S_{\mathfrak{S}}} \log\left(1 + e^{-f_{cal}(s;\mathbf{\Pi}) - \log\frac{\pi}{1-\pi}}\right) \\ & + \frac{1-\pi}{|S_{\mathfrak{D}}|} \sum_{s \in S_{\mathfrak{D}}} \log\left(1 + e^{f_{cal}(s;\mathbf{\Pi}) + \log\frac{\pi}{1-\pi}}\right) \end{aligned} \tag{45}$$

where π is the effective prior for the target class. As we show in the experimental section, discriminative training obtains similar results as generative training, confirming that

our model can provide a good characterization of the score distributions.

V. Utterance-dependent calibration

According to our model, a score for a trial $\mathbf{z} = \left[\boldsymbol{\phi}_{\mathcal{E}}^T, \boldsymbol{\phi}_{\mathcal{T}}^T\right]^T$ can be interpreted as a sample of a R.V. with conditional VΓ-distributions given by (40), 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 to utterance. In particular, for each speaker-vector $\boldsymbol{\phi}_i$ we let the effective within-class variance w_i be a function of both i.i.d. and utterance-dependent miscalibration sources.

A. Utterance duration

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 [22] models. The i-vector model allows accounting for i-vector uncertainty through the i-vector posterior covariance matrix [28]–[30]. Incorporating the i-vector uncertainty at trial level is equivalent to modeling target and non-target trials as in (13), 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 [35] by a matrix whose form is

$$\mathbf{C} \approx (\mathbf{I} + D\mathbf{M})^{-1} \ , \tag{46}$$

where $\mathbf{M}$ depends on the UBM and the i-vector model parameters, whereas D is the utterance duration. We further assume that $\mathbf{C}$ has the same principal directions as $\mathbf{W}_{\mathcal{M}}$, $\mathbf{W}_{\mathcal{E}}$ and $\mathbf{W}_{\mathcal{T}}$, so that they can be jointly diagonalized (a similar approximation was used in [36]). Each component $\mathbf{C}_j$ of $\mathbf{C}$ has then a functional form $\mathbf{C}_j = \frac{1}{1+D\eta_j^{-1}} = \frac{\eta_j}{\eta_j + D}$. 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 $w_{\mathcal{T},i}$ for a trial $\mathbf{z}_i$ as

$$w_{\mathcal{E},i} = w_{\mathcal{E}} + \frac{\psi}{D_{\mathcal{E},i} + \eta} \ , \quad w_{\mathcal{T},i} = w_{\mathcal{T}} + \frac{\psi}{D_{\mathcal{T},i} + \eta} \ , \tag{47}$$

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. As for the VΓ-Var model, also in this case we can assume 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 [20]), or uncertainty cannot be taken into account at classification level (e.g. PSVM).

B. Speaker vector norms as a proxy for utterance-dependent variability

Standard PLDA models employ length normalization to improve the classification accuracy. In this section we show that the squared norm of a speaker vector can, indeed, be interpreted as a measure of utterance-dependent variability, and length normalization as a way to reduce intra-trial mismatch. Our derivations are based on the model of [37] that introduces a simplified Heavy-Tailed PLDA (HT-PLDA):

$$\begin{gathered} \boldsymbol{\Phi}_i = \mathbf{m} + \overline{\mathbf{U}}\mathbf{Y} + \mathbf{E}_i \ , \\ \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ , \quad \mathbf{E}_i | V_i \sim \mathcal{N}(\mathbf{0}, V_i \mathbf{W}) \ , \quad V_i^{-1} \sim \Gamma\left(\frac{a}{2}, \frac{a}{2}\right) \end{gathered} \tag{48}$$

where V_i is a hidden R.V. representing an utterance-dependent scaling factor. The HT-PLDA model assumes a gamma prior for the inverse of V_i. In typical PLDA applications the speaker subspace is of much smaller dimension than the speaker vector space, thus we also assume that the dimensionality of $\mathbf{Y}$ is $D \ll M$. Without loss of generality, we also assume that speaker vectors have been whitened so that $\mathbf{m} = \mathbf{0}$, $\mathbf{W} = \mathbf{I}$ and the D-dimensional speaker subspace corresponds to the first D principal directions of the speaker vector space. The speaker vector can be partitioned in two components:

$$\overline{\mathbf{U}} = \begin{bmatrix} \mathbf{U} \\ \mathbf{0} \end{bmatrix} \ , \quad \boldsymbol{\Phi}_i = \begin{bmatrix} \boldsymbol{\Phi}_i^u \\ \boldsymbol{\Phi}_i^n \end{bmatrix} = \begin{bmatrix} \mathbf{U}\mathbf{Y} + \mathbf{E}_i^u \\ \mathbf{E}_i^n \end{bmatrix} \tag{49}$$

where $\mathbf{U}$ is a $D \times D$ matrix, $\boldsymbol{\Phi}_i^u$ refers to the first d components of $\boldsymbol{\Phi}_i$, $\boldsymbol{\Phi}_i^n$ refers to the remaining components of $\boldsymbol{\Phi}_i$, which do not depend on the speaker factor $\mathbf{y}$, and similarly for the noise terms $\mathbf{E}_i^u$, $\mathbf{E}_i^n$.The marginal density for a set of k vectors belonging to a single speaker is

$$\begin{aligned} & f_{\Phi_1 \dots \Phi_k}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_k) = \\ & = \int f_{\boldsymbol{\Phi}_1 \dots \boldsymbol{\Phi}_k | \mathbf{Y}, V_1 \dots V_k}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_k | \mathbf{y}, v_1 \dots v_k) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \prod_{i=1}^{k} f_{V_i}(v_i) dv_i \\ & = \int \prod_{i=1}^{k} f_{\boldsymbol{\Phi}_i^u | \mathbf{Y}, V_i}\left(\boldsymbol{\phi}_i^u | \mathbf{y}, v_i\right) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \prod_{i=1}^{k} f_{\boldsymbol{\Phi}_i^n | V_i}\left(\boldsymbol{\phi}_i^n | v_i\right) f_{V_i}(v_i) dv_i \\ & = \prod_{i=1}^{k} f_{\boldsymbol{\Phi}_i^n}\left(\boldsymbol{\phi}_i^n\right) \int \prod_{i=1}^{k} f_{\boldsymbol{\Phi}_i^u | \mathbf{Y}, V_i}\left(\boldsymbol{\phi}_i^u | \mathbf{y}, v_i\right) f_{V_i | \boldsymbol{\Phi}_i^n}\left(v_i | \boldsymbol{\phi}_i^n\right) dv_i f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}. \end{aligned} \tag{50}$$

The LLR for a given trial depends only on the integral term of equation (50). We can observe that $V_i | \boldsymbol{\Phi}_i^n$ acts as a prior for the conditional likelihood of $\boldsymbol{\Phi}_i^u | \mathbf{Y}, V_i$, with

$$\left(V_i^{-1} | \boldsymbol{\Phi}_i^n = \boldsymbol{\phi}_i^n\right) \sim \Gamma\left(\frac{a + M - D}{2}, \frac{a + \|\boldsymbol{\phi}_i^n\|^2}{2}\right) \tag{51}$$

Rather than working with the posterior distribution, we consider an approximation that replaces the posterior distribution for V_i with a Maximum-a-Posteriori (MAP) point estimate[6]:

$$v_i^{\text{MAP}} = \frac{a + \|\boldsymbol{\phi}_i^n\|^2}{a + M - D + 1} \ . \tag{52}$$

[6]The posterior distribution of $V_i | \boldsymbol{\Phi}_i^n = \boldsymbol{\phi}_i^n$ is an inverse-Gamma with the same parameters of (51). The MAP estimate for V_i^{-1} is slightly different, but has a similar formal expression.

Replacing R.V. V_i with the MAP estimate (52) we obtain again a standard PLDA model, where the noise variance for embedding i corresponds to $v_i^{\text{MAP}}\mathbf{I}$. It is worth noting that, rather than estimating the HT-PLDA model, we can approximate matrix $\overline{\mathbf{U}}$ with the subspace found by Linear Discriminant Analysis. The MAP solution would then correspond to the norm of the embedding computed in the complement of the LDA subspace. If we assume that the speaker subspace is small-dimensional, and that the between-to-within variability ratio is nevertheless small for most embedding directions, then the MAP solution can be further approximated by the norm of the original, whitened embedding. Length normalization has thus the effect of making the noise term distribution approximately utterance-independent, although it introduces a dependency of the speaker factor distribution on the utterance. Since the estimated $\mathbf{v}_i^{\text{MAP}}$ acts as a proxy of utterance-dependent variability, it can alternatively be incorporated in our calibration model by considering effective enrollment and test variances, defined as

$$w_{\mathcal{E},i} = w_{\mathcal{C}} + \psi \left\| \mathbf{U}_c \boldsymbol{\phi}_{\mathcal{E},i} \right\|^2, \; w_{\mathcal{T},i} = w_{\mathcal{C}} + \psi \left\| \mathbf{U}_c \boldsymbol{\phi}_{\mathcal{T},i} \right\|^2. \quad (53)$$

where $\mathbf{U}_c$ is a projection matrix corresponding to the complement of an LDA subspace, and ψ is a parameter that can be estimated by Maximum Likelihood over the calibration set.

VI. Experiments

In this section we analyze the performance of the proposed models with different speaker vector front-ends and classification back-ends on different test sets.

A. Classification back-ends

We consider two different back-ends, PLDA and Pairwise Support Vector Machine (PSVM) [23], [24]. The PSVM approach trains a single classifier on speaker vector pairs aimed at separating same-speaker from different-speaker trials. The separation surfaces are derived from the PLDA log-likelihood ratio expression, whereas the model is trained using the standard SVM objective.

B. Evaluation sets, front-ends and training data

We consider three datasets, SRE 2019[7], SRE 2012[8] and SRE 2010[9], with different front-ends.

1) SRE 2019: For the SRE 2019 dataset we consider three embedding extractors, trained on a common list including VoxCeleb1 and VoxCeleb2 [38], Mixer 4,5 and 6 and Switchboard data[10]. The MUSAN [39] and the AIR [40] datasets were used for data augmentation. The first extractor is based on a Time-Delay Neural Network (TDNN) with the same topology as in [41]. The DNN input consists of 24-dimensional Perceptual Linear Predictors (PLP) features, and the speaker embeddings are 512-dimensional.[11]. The embeddings have been further processed by means of Linear Discriminant Analysis (LDA), which reduces the dimensionality to 400 for PSVM and 200 for PLDA. The second front-end is based on a Factorized Time-Delay Neural Network (FTDNN) [42], implemented as in [43], trained using softmax and cross-entropy loss on clean data only. Embeddings are 512-dimensional. As for TDNN, LDA was applied to reduce the embedding dimensionality. The third extractor is based on the ECAPA architecture [44]. The network has been trained using Additive Angular Margin softmax [45] and cross-entropy loss. The embeddings are 192-dimensional. In this case no further dimensionality reduction was applied. For both back-ends we analyzed the effects of an additional length normalization step applied to whitened embeddings. The PSVM pre-processing pipeline further includes a Within-Class Covariance (WCCN) normalization step, applied either on the raw embeddings, or on the length-normalized vectors. PLDA has been trained on SRE 2018 Evaluation data, comprising about 13 thousand segments. Since PSVM requires more data to provide good results, PSVM models were trained with an additional subset of Mixer 4, 5 and 6, Switchboard and VoxCeleb data, for a total of about 110 thousand segments. The training lists were chosen based on the best results of each baseline model on SRE 2019 Progress data. The calibration models were trained on a subset of the SRE 2019 Progress set. The calibration performance was evaluated on the SRE 2019 Evaluation set.

2) SRE 2012: The SRE 2012 system is based on a hybrid GMM/DNN model [46], [47] trained using SRE-04 to SRE-10 and Switchboard data, for a total of about 42 thousand segments. The acoustic features consist of 20 PLP coefficients and their delta and delta-delta parameters. The DNN comprises 256 outputs. For each DNN output, we fit an 8-dimensional, full covariance GMM using the approach in [48]. Overall, the UBM has 2048 components. The speaker vectors are obtained from a 400-dimensional e-vector extractor [27]. The back-ends have been trained using the same front-end datasets. Tests were performed on the extended tel-tel core condition (condition 5). The test set was divided into two, non overlapping parts. The first part, which comprises 25% of the enrollment segments, was used to estimate the calibration parameters. The remaining part was used as evaluation.

3) SRE 2010: The SRE 2010 system is based on 400-dimensional i-vectors, estimated from a gender-dependent, 1024-components, diagonal covariance UBM based on 45-dimensional MFCC features, incorporating delta and double-delta parameters. The back-end is a PLDA classifier. I-vectors pre-processing consists of whitening and length normalization. The front-end has been trained using Switchboard, SRE-04 to SRE-06 and Fisher data. The back-end has been trained on the same lists, but without the Fisher dataset. Tests were performed on the female extended tel-tel condition (condition 5). The test set was divided into two, non overlapping, parts. The first part, comprising 25% of the enrollment segments, was used to estimate the calibration parameters. The remaining part was used for evaluation.

For all test sets, both calibration and evaluation embeddings have been centered with respect to the calibration set

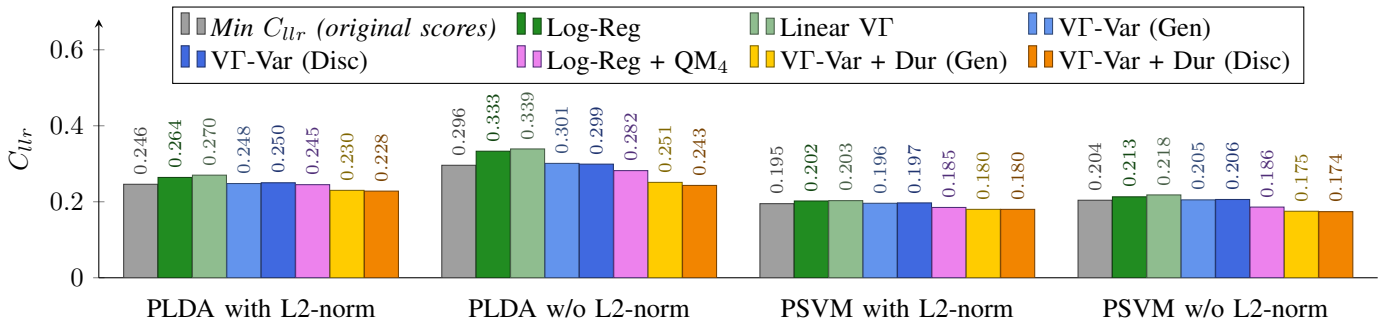
[7]https://www.nist.gov/publications/2019-nist-speaker-recognition-evaluation-cts-challenge

[8]https://www.nist.gov/itl/iad/mig/sre12-results

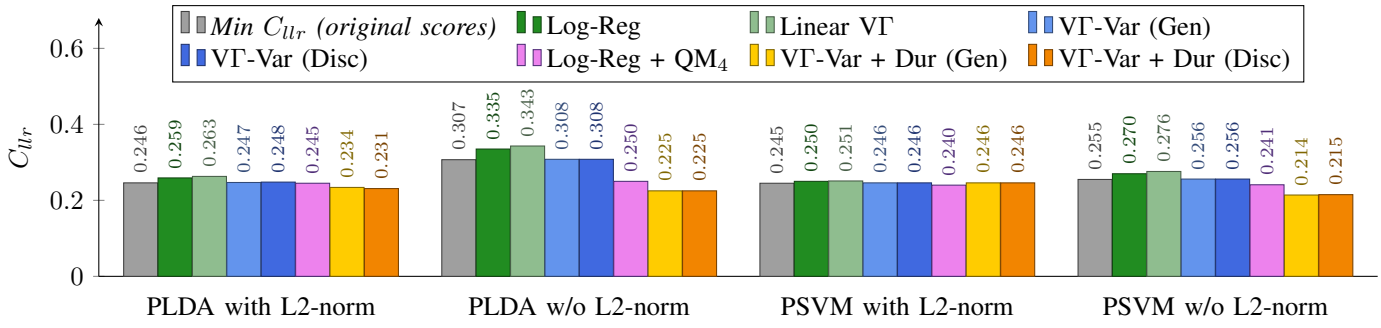
[9]https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2010

[10]Available from Linguistic Data Consortium

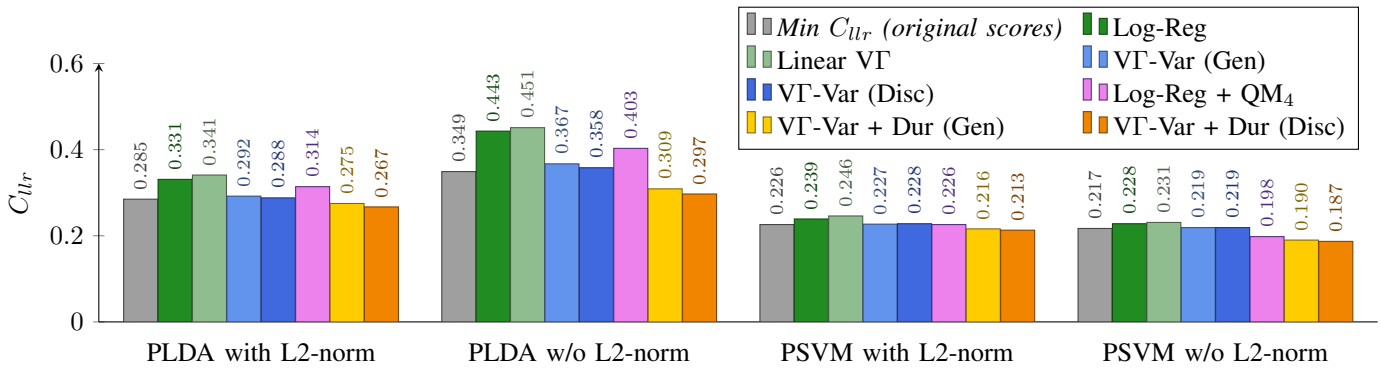
[11]Since we employed a different training set, the results are not directly comparable to those we previously published in [21] and [14] for the Linear VG model



(a) ECAPA embeddings



(b) FTDNN embeddings



(c) TDNN embeddings

Fig. 1: C_{llr} for different calibration approaches with different embedding front-ends on SRE 2019 short segments. Calibration models have been trained with target prior $\pi = 0.1$ (discriminative models) or target weight $\zeta = 0.1$ (generative models).

embedding mean, to account for possibly non-zero evaluation population means (cf. Section III, where we assumed that the evaluation populations have zero-mean distribution, i.e. $\mathbf{m} = \mathbf{0}$ in (14)).

C. Results on SRE 2019

The first set of experiments shows the results of the VΓ-Var and VΓ-Var + Dur models on SRE 2019 data. Since duration models are more effective in presence of short and variable utterance duration, we first report results on short cuts obtained from the original SRE 2019 data. In particular, both enrollment and test segments have been randomly truncated to a length between 3 and 30 seconds. In Fig. 1 and Table I we compare the performance of different calibration approaches for different front-end and back-end combinations in terms of C_{llr} [4], [49], [50]. In Table I we additionally report the actual primary cost C_{prim} as defined by NIST for the SRE 2019 evaluation, and, since duration modeling improves discrimination, we also provide Equal Error Rates (EER). The first row of each table reports minimum costs computed on the original classifier outputs. The following four rows show the performance of calibration models that do not employ side-information. The baseline systems consist of a prior-weighted Logistic Regression model (Log-Reg) and the linear, Constrained ML Variance-Gamma model of [14] (Linear VΓ). Rows labeled VΓ-Var (Gen) and VΓ-Var (Disc) report the

TABLE I: Results on the SRE 2019 evaluation dataset with short segments. Calibration models have been trained with target prior $\pi = 0.1$ (discriminative models) or target weight $\zeta = 0.1$ (generative models).

(a) ECAPA embeddings

	PLDA						PSVM					
	with L2-norm			w/o L2-norm			with L2-norm			w/o L2-norm		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs†	*0.246*	*0.485*	*6.8%*	*0.296*	*0.486*	*8.7%*	*0.195*	*0.448*	*5.3%*	*0.204*	*0.452*	*5.6%*
Log-Reg	0.264	0.495	6.8%	0.333	0.571	8.7%	0.202	0.457	5.3%	0.213	0.466	5.6%
Linear VΓ	0.270	0.488	6.8%	0.339	0.551	8.7%	0.203	0.456	5.3%	0.218	0.458	5.6%
VΓ-Var (Gen)	0.248	0.486	6.8%	0.301	0.487	8.7%	0.196	0.449	5.3%	0.205	0.454	5.6%
VΓ-Var (Disc)	0.250	0.485	6.8%	0.299	0.487	8.7%	0.197	0.449	5.3%	0.206	0.454	5.6%
Log-Reg + QM_4	0.245	0.493	6.1%	0.282	0.532	6.7%	0.185	0.467	4.8%	0.186	0.484	4.7%
VΓ-Var + Dur (Gen)	0.230	0.485	6.1%	0.251	0.484	6.6%	0.180	0.447	4.8%	0.175	0.454	4.6%
VΓ-Var + Dur (Disc)	0.228	0.478	6.1%	0.243	0.461	6.6%	0.180	0.448	4.8%	0.174	0.454	4.6%

(b) FTDNN embeddings

	PLDA						PSVM					
	with L2-norm			w/o L2-norm			with L2-norm			w/o L2-norm		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs†	*0.246*	*0.466*	*7.0%*	*0.307*	*0.477*	*9.0%*	*0.245*	*0.536*	*7.0%*	*0.255*	*0.529*	*7.1%*
Log-Reg	0.259	0.482	7.0%	0.335	0.564	9.0%	0.250	0.550	7.0%	0.270	0.553	7.1%
Linear VΓ	0.263	0.474	7.0%	0.343	0.531	9.0%	0.251	0.544	7.0%	0.276	0.534	7.1%
VΓ-Var (Gen)	0.247	0.467	7.0%	0.308	0.477	9.0%	0.246	0.537	7.0%	0.256	0.531	7.1%
VΓ-Var (Disc)	0.248	0.466	7.0%	0.308	0.477	9.0%	0.246	0.537	7.0%	0.256	0.530	7.1%
Log-Reg + QM_4	0.245	0.486	6.5%	0.250	0.520	6.4%	0.240	0.558	6.6%	0.241	0.591	6.3%
VΓ-Var + Dur (Gen)	0.234	0.468	6.4%	0.225	0.457	6.1%	0.246	0.537	7.0%	0.214	0.518	5.9%
VΓ-Var + Dur (Disc)	0.231	0.465	6.4%	0.225	0.454	6.1%	0.246	0.537	7.0%	0.215	0.520	5.9%

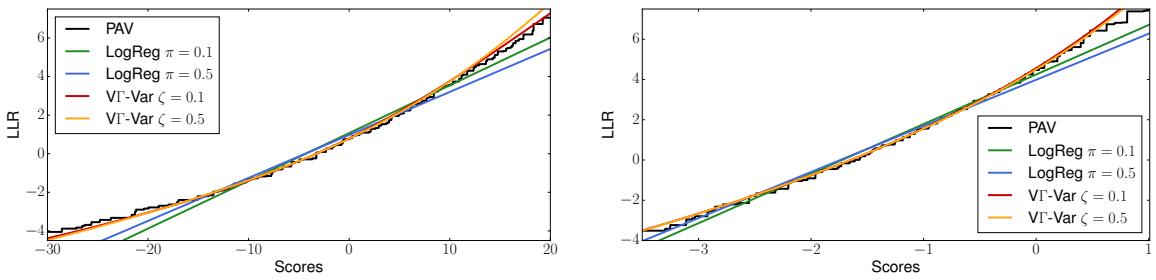
(c) TDNN embeddings

	PLDA						PSVM					
	with L2-norm			w/o L2-norm			with L2-norm			w/o L2-norm		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs†	*0.285*	*0.514*	*7.9%*	*0.349*	*0.532*	*10.2%*	*0.226*	*0.480*	*6.1%*	*0.217*	*0.462*	*5.9%*
Log-Reg	0.331	0.528	7.9%	0.443	0.652	10.2%	0.239	0.486	6.1%	0.228	0.476	5.9%
Linear VΓ	0.341	0.517	7.9%	0.451	0.623	10.2%	0.246	0.480	6.1%	0.231	0.464	5.9%
VΓ-Var (Gen)	0.292	0.526	7.9%	0.367	0.536	10.2%	0.227	0.480	6.1%	0.219	0.463	5.9%
VΓ-Var (Disc)	0.288	0.526	7.9%	0.358	0.537	10.2%	0.228	0.481	6.1%	0.219	0.464	5.9%
Log-Reg + QM_4	0.314	0.529	7.2%	0.403	0.613	8.3%	0.226	0.488	5.7%	0.198	0.511	5.0%
VΓ-Var + Dur (Gen)	0.275	0.537	7.2%	0.309	0.536	8.3%	0.216	0.487	5.7%	0.190	0.465	4.9%
VΓ-Var + Dur (Disc)	0.267	0.536	7.1%	0.297	0.534	8.2%	0.213	0.479	5.6%	0.187	0.472	5.0%

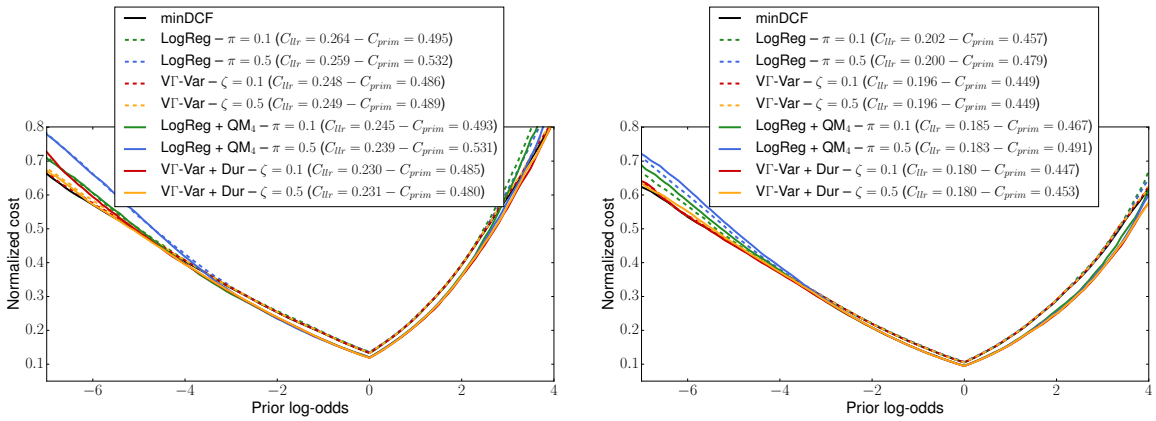
† Minimum C_{llr}, minimum C_{prim}, and EER computed on the classifier scores. Models using side-information may provide lower costs.

results of our proposed VΓ-Var model (42) trained with a generative, ML criterion (43) and a discriminative objective (45), respectively. For generative models we set the target weight to $\zeta = 0.1$. For discriminative models the target prior was set to $\pi = 0.1$. The last three rows show the performance of calibration approaches that incorporate duration information. Our baseline follows the approach of [8], which enriches the linear Log-Reg calibration model with quality measures 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. For each back-end, we report both results without (blue columns) and with (yellow columns) embedding length normalization. For the ECAPA front-end we also show in Fig. 2-a and 2-b the calibration transformations and Bayes error plots for the duration-agnostic models, and the Bayes Error plots for duration-aware models, trained with different target priors π or target weights ζ.

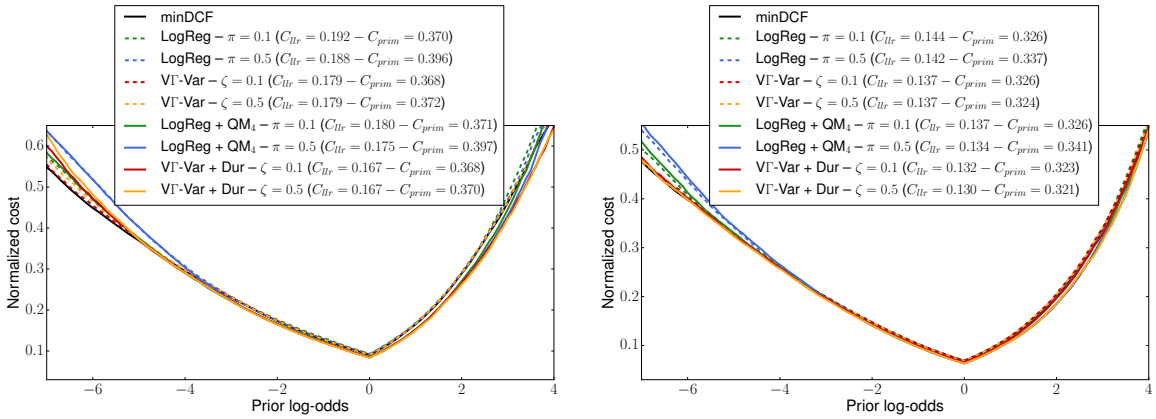
Concerning the baseline systems, the minimum cost results show that the PSVM back-end significantly outperforms PLDA, both with and without length normalization. Further-



(a) Calibration transformation for short segments - duration-agnostic models - left: PLDA, right: PSVM



(b) Bayes error plot for short segments - left: PLDA, right: PSVM



(c) Bayes error plot for original segments - left: PLDA, right: PSVM

Fig. 2: Calibration transformation and Bayes error plots for different calibration models and different back-ends on SRE 2019 evaluation data, ECAPA front-end.

more, we can observe that length normalization allows for a significant reduction of C_{llr} and EER for PLDA, whereas for PSVM the benefits of the normalization are significantly smaller, and for TDNN it actually decreases performance. Concerning calibration, Logistic Regression and Linear VΓ provide similar results, the former achieving slightly lower C_{llr} at the cost of a degradation in terms of C_{prim}. Except for the TDNN-PLDA system, where both models provide significant miscalibration error, these approaches produce reasonably good results for length-normalized embeddings, but incur in a significant degradation for PLDA models trained over non-normalized embeddings. Fig. 2-a compares the calibration transformations of the different models with the solution provided by isotonic regression computed through the PAV [51] algorithm on the calibration set for the ECAPA front-end with length-normalized embeddings. We can observe that a linear approximation of the PAV solution cannot provide optimal calibration over the whole range of scores. Without length normalization (not shown in the figures due to lack of space), the non-linearity becomes even more evident, and results in a significant calibration loss for linear models. The proposed VΓ-Var approach, on the contrary, intrinsically computes non-linear mappings that account for distribution mismatches, and is therefore able to provide a better fit to the PAV calibration

TABLE II: Results on the SRE 2019 evaluation dataset with original segments. . Calibration models have been trained with target prior $\pi = 0.1$ (discriminative models) or target weight $\zeta = 0.1$ (generative models).

(a) ECAPA embeddings

	PLDA						PSVM					
	with L2-norm			w/o L2-norm			with L2-norm			w/o L2-norm		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs†	*0.176*	*0.365*	*4.6%*	*0.204*	*0.358*	*5.6%*	*0.135*	*0.323*	*3.5%*	*0.135*	*0.330*	*3.5%*
Log-Reg	0.192	0.370	4.6%	0.237	0.405	5.6%	0.144	0.326	3.5%	0.146	0.334	3.5%
Linear VΓ	0.200	0.366	4.6%	0.245	0.388	5.6%	0.147	0.326	3.5%	0.151	0.331	3.5%
VΓ-Var (Gen)	0.179	0.368	4.6%	0.210	0.360	5.6%	0.137	0.326	3.5%	0.137	0.334	3.5%
VΓ-Var (Disc)	0.179	0.367	4.6%	0.206	0.360	5.6%	0.139	0.326	3.5%	0.137	0.335	3.5%
Log-Reg + QM_4	0.180	0.371	4.3%	0.217	0.395	5.0%	0.137	0.326	3.2%	0.137	0.340	3.1%
VΓ-Var + Dur (Gen)	0.167	0.368	4.3%	0.190	0.356	5.0%	0.132	0.323	3.3%	0.128	0.340	3.1%
VΓ-Var + Dur (Disc)	0.165	0.362	4.2%	0.186	0.350	5.0%	0.130	0.324	3.2%	0.127	0.334	3.1%

(b) FTDNN embeddings

	PLDA						PSVM					
	with L2-norm			w/o L2-norm			with L2-norm			w/o L2-norm		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs†	*0.163*	*0.326*	*4.3%*	*0.185*	*0.329*	*5.0%*	*0.150*	*0.366*	*3.9%*	*0.145*	*0.350*	*3.7%*
Log-Reg	0.176	0.331	4.3%	0.202	0.354	5.0%	0.158	0.370	3.9%	0.155	0.358	3.7%
Linear VΓ	0.182	0.327	4.3%	0.212	0.334	5.0%	0.160	0.367	3.9%	0.161	0.352	3.7%
VΓ-Var (Gen)	0.165	0.330	4.3%	0.186	0.330	5.0%	0.152	0.367	3.9%	0.147	0.351	3.7%
VΓ-Var (Disc)	0.165	0.329	4.3%	0.186	0.331	5.0%	0.152	0.367	3.9%	0.146	0.353	3.7%
Log-Reg + QM_4	0.165	0.330	4.0%	0.172	0.349	4.3%	0.150	0.376	3.6%	0.144	0.374	3.4%
VΓ-Var + Dur (Gen)	0.155	0.332	4.0%	0.160	0.326	4.2%	0.146	0.372	3.7%	0.134	0.354	3.3%
VΓ-Var + Dur (Disc)	0.153	0.329	3.9%	0.157	0.316	4.2%	0.143	0.370	3.6%	0.133	0.353	3.3%

(c) TDNN embeddings

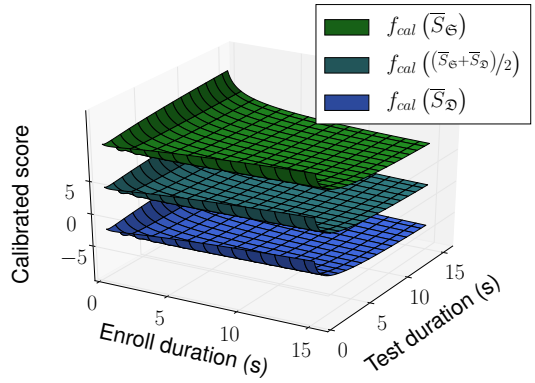
	PLDA						PSVM					
	with L2-norm			w/o L2-norm			with L2-norm			w/o L2-norm		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs†	*0.200*	*0.387*	*5.2%*	*0.236*	*0.396*	*6.4%*	*0.164*	*0.366*	*4.2%*	*0.140*	*0.339*	*3.6%*
Log-Reg	0.231	0.388	5.2%	0.287	0.438	6.4%	0.177	0.370	4.2%	0.150	0.340	3.6%
Linear VΓ	0.246	0.392	5.2%	0.301	0.412	6.4%	0.189	0.370	4.2%	0.154	0.341	3.6%
VΓ-Var (Gen)	0.206	0.407	5.2%	0.244	0.403	6.4%	0.165	0.366	4.2%	0.142	0.341	3.6%
VΓ-Var (Disc)	0.204	0.403	5.2%	0.238	0.406	6.4%	0.165	0.367	4.2%	0.141	0.344	3.6%
Log-Reg + QM_4	0.218	0.391	4.8%	0.267	0.435	5.8%	0.169	0.366	4.0%	0.140	0.349	3.2%
VΓ-Var + Dur (Gen)	0.195	0.420	4.9%	0.223	0.409	5.8%	0.160	0.381	4.0%	0.133	0.351	3.2%
VΓ-Var + Dur (Disc)	0.191	0.431	4.8%	0.215	0.398	5.7%	0.155	0.363	3.9%	0.130	0.345	3.2%

† Minimum C_{llr}, minimum C_{prim}, and EER computed on the classifier scores. Models using side-information may provide lower costs.

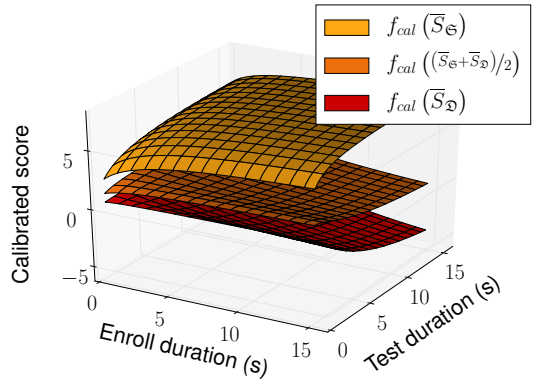
transformation, obtaining actual costs that are very close to the minimum ones both for PLDA and PSVM, with and without length normalization. The generative and discriminative VΓ-Var models obtain similar results, suggesting that VΓ-Var provides an accurate characterization of target and non-target score distributions. Fig. 2-b shows the effects of training with different target prior π or target weight ζ for the ECAPA front-end (duration-agnostic models, dashed lines). We can observe that Logistic Regression is sensitive to the chosen prior — setting $\pi = 0.5$ results in a slightly lower C_{llr}, however it significantly increases the actual primary cost C_{prim}. The VΓ-Var models, on the other hand, are less affected by the choice of ζ, providing similar results with both $\zeta = 0.1$ and $\zeta = 0.5$. Concerning duration-aware models, both Log-Reg + QM_4 and the proposed VΓ-Var + Dur outperform the corresponding duration-agnostic methods both in terms of C_{llr} and EER for almost all front-end and back-end combinations, showing both better calibration and discrimination capabilities. The VΓ-Var + Dur approach consistently outperforms quality measure based methods in terms of C_{llr}, with relative improvements

TABLE III: Comparison of length-normalization-aware calibration models on the SRE 2019 evaluation dataset.

		Short segments						Original segments					
		PLDA			PSVM			PLDA			PSVM		
L2-norm	Calibration	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
ECAPA													
✓	VΓ-Var (Gen)	0.248	0.486	6.8%	0.196	0.449	5.3%	0.179	0.368	4.6%	0.137	0.326	3.5%
✗	VΓ-Var + Norm (Gen)	0.229	0.483	5.9%	0.188	0.453	4.9%	0.175	0.363	4.3%	0.135	0.329	3.3%
FTDNN													
✓	VΓ-Var (Gen)	0.247	0.467	7.0%	0.246	0.537	7.0%	0.165	0.330	4.3%	0.152	0.367	3.9%
✗	VΓ-Var + Norm (Gen)	0.222	0.457	6.1%	0.215	0.510	5.9%	0.160	0.319	4.1%	0.135	0.345	3.4%
TDNN													
✓	VΓ-Var (Gen)	0.292	0.526	7.9%	0.227	0.480	6.1%	0.206	0.407	5.2%	0.165	0.366	4.2%
✗	VΓ-Var + Norm (Gen)	0.275	0.536	6.9%	0.198	0.464	5.1%	0.211	0.406	5.1%	0.141	0.337	3.4%



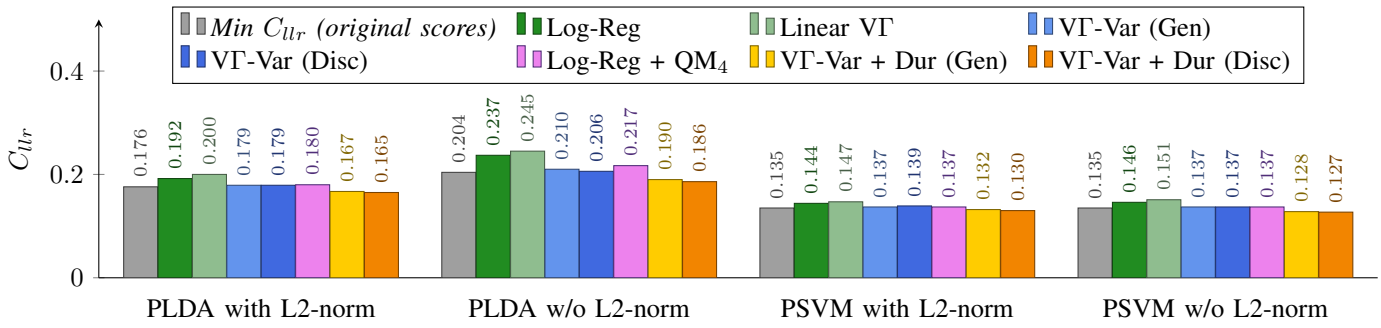
(a) Log-Reg + QM_4



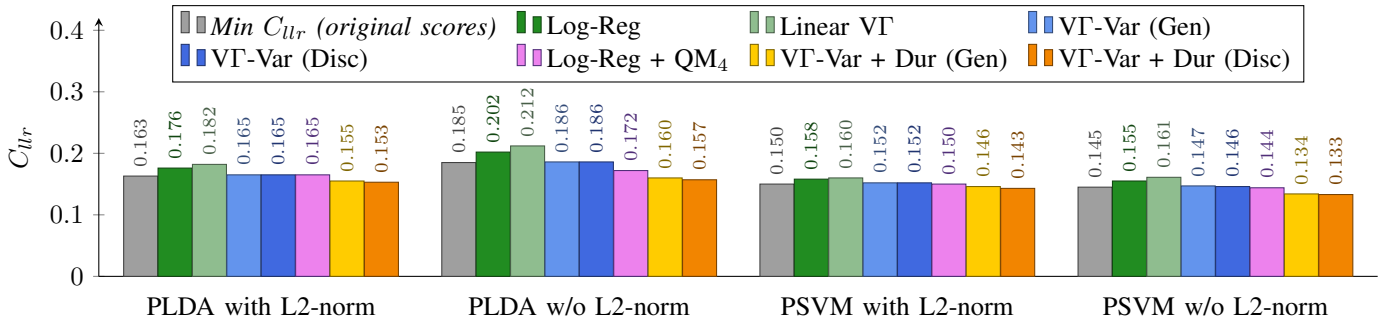
(b) VΓ-Var + Dur (Gen)

Fig. 3: Calibration transformations of duration-aware models for different, fixed scores as a function of segments duration. ECAPA embeddings without length normalization, PLDA back-end.

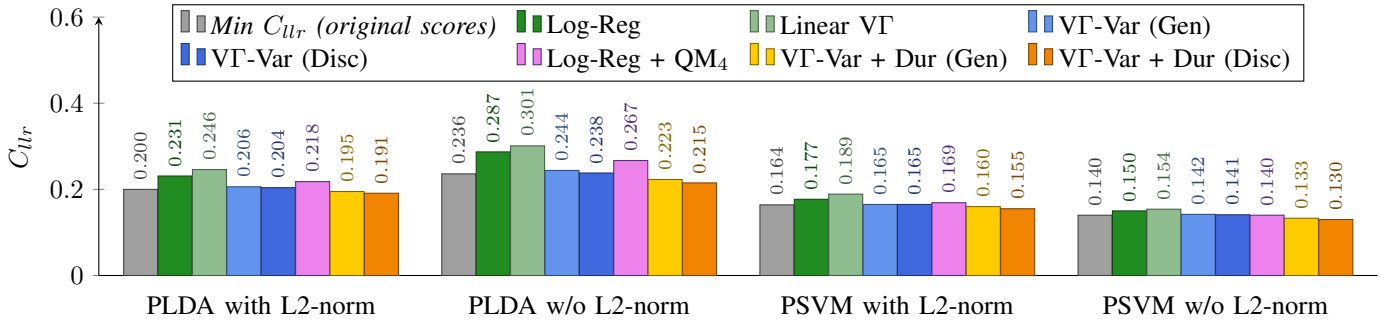
ranging from 3% to 15% for length normalized embeddings, and from 6% to 26% for non-normalized embeddings, while providing similar or slightly better EERs. We observe a small, 2% relative degradation with respect to Log-Reg + QM_4 only for the FTDNN / PSVM combination with length-normalized embeddings. In this case, however, we can observe that the best results are still obtained by combining the proposed VΓ-Var + Dur approach with non-normalized embeddings. Concerning C_{prim} costs, Log-Reg + QM_4 results are not consistent, with the model providing similar or slightly worse results with respect standard Log-Reg for back-ends that employ length normalization. For back-ends trained over non-normalized embeddings we see an improvement for PLDA but a significant degradation for PSVM models. Compared to Log-Reg + QM_4, our approach provides better C_{prim}, although on average the C_{prim} is close to that of our VΓ-Var model. Finally, Fig. 2-b shows the Bayes error plots of Log-Reg + QM_4 and generative VΓ-Var + Dur models (duration-aware models, solid lines). Also in this case, we observe that Log-Reg + QM_4 is sensitive to the choice of the training prior, whereas our approach provides consistent results for both $\zeta = 0.1$ and $\zeta = 0.5$. To better understand the differences between Log-Reg + QM_4 and VΓ-Var + Dur models we show, in Figure 3, how the same score, produced by different trials, would be transformed by the two models as a function of the trial enrollment and test duration. For each model we consider 3 scores, namely the average target score $\overline{S}_{\mathfrak{S}}$, the average non-target score $\overline{S}_{\mathfrak{D}}$ and their mean $\frac{\overline{S}_{\mathfrak{S}}+\overline{S}_{\mathfrak{D}}}{2}$, corresponding to the three surfaces shown in the figure. The plot refers to the ECAPA-PLDA with non-normalized embeddings system. This system was chosen because the effects are graphically more pronounced. Similar considerations, however, hold also for the other systems. We can observe that Log-Reg + QM_4 models provide the same kind of transformation regardless of the score value. Indeed, quality measures simply act as an additive term to the score. On the contrary, VΓ-Var models provide transformations that affect differently target and non-target scores. We can observe that shorter duration results in scores being driven toward a LLR of zero. This is consistent with the fact that shorter segments have naturally larger uncertainty, which is not taken into account at the classification level [28]. The back-end scores tend to be over-confident for shorter utterances, and the VΓ-Var model is compensating this behaviour.



(a) ECAPA embeddings



(b) FTDNN embeddings



(c) TDNN embeddings

Fig. 4: C_{llr} for different calibration approaches with different embedding front-ends on SRE 2019 original segments. Calibration models have been trained with target prior $\pi = 0.1$ (discriminative models) or target weight $\zeta = 0.1$ (generative models).

It is worth noting that for all front-ends duration models are less effective in presence of length-normalized embeddings. This is consistent with our assumption that length normalization is actually normalizing the within-class variance of the embeddings, thus reducing also variability due to duration. Since duration variability has been partially compensated, duration-aware calibration models are less able to further improve accuracy. We can observe that, at least for PSVM, it's actually more effective accounting for duration variability at calibration stage, rather than relying on length normalization.

For completeness, we report in Fig. 4, Table II and Fig. 2-c the results of the different calibration approaches for the original SRE 2019 segments. Similar considerations as for short segments apply, the main difference being that, as expected, duration-aware models provide smaller relative improvements.

A second set of experiments analyzes the effectiveness of accounting for the embedding norm at calibration rather than at classification level. Table III compares VΓ-Var calibration for length-normalized back-ends with the normalization-aware VΓ-Var + Norm model (53) paired with back-ends that do not employ length normalization. In both cases, the embedding norm was computed from the whole vector. The yellow rows refer to normalized embeddings with VΓ-Var calibration, whereas the blue rows refer to raw embeddings and VΓ-Var + Norm models. Although standard PLDA with standard calibration benefits from length normalization, we can observe that incorporating the embedding norm at calibration level rather than at scoring level is, for this dataset, actually

beneficial. The VΓ-Var + Norm model, combined with PLDA scores from non-normalized embeddings, provides significantly lower C_{llr} and EER for short segments, with no impact on actual C_{prim} costs. For longer segments we do not observe a significant improvement over length-normalized back-ends, although the VΓ-Var + Norm model provides slightly lower EER and similar performance in terms of C_{llr} and C_{prim}. For PSVM we observe, on average, a significant reduction of C_{llr} and EER when the embedding norm is accounted for at calibration level. For FTDNN and TDNN embeddings we also obtain an improvement in terms of C_{prim}, whereas for ECAPA embeddings the results of the two approaches are very similar. It is worth noting that combining VΓ-Var + Norm with length-normalized embeddings would not provide additional benefits over VΓ-Var models. Indeed, the norm that should be employed at calibration level is the norm of the input vectors of the PLDA model, i.e. the norm of the length-normalized vectors. Since this norm would be the same for all embeddings, the VΓ-Var and VΓ-Var + Norm models would become equivalent.

Finally, we also considered a simple model that combines both duration and embedding norm variability as

$$w_{\mathcal{S},i} = w_{\mathcal{C}} + w_{dur,i} + w_{norm,i} \tag{54}$$

where $w_{dur,i}$ and $w_{norm,i}$ are the duration and norm components used in VΓ-Var + Dur and VΓ-Var + Norm models. For SRE 2019, we observed that incorporating both terms does not provide additional benefits, suggesting that both terms convey similar side information.

D. SRE 2010

The SRE 2010 test contains mainly long utterances. As for SRE 2019, also in this case we consider a modified test set where segments have been cut to a duration between 3 and 60 seconds. The i-vector front-end allows quantifying the uncertainty in the i-vector estimate, which is strongly related to utterance duration. For the short segments setup we therefore consider as additional baseline a Full-Posterior-Distribution PLDA (FPD-PLDA) model [28]. The results are reported in Table IV. The first row shows the minimum costs of PLDA and FPD-PLDA for short segments, and the minimum costs of the same PLDA model for the original segments. For short segments we can observe that FPD-PLDA significantly improves both C_{llr} and EER, while the actual C_{prim} costs of the two models are very close. The next four rows compare the duration-agnostic calibration models. As for SRE 2019, the VΓ-Var models provide the lowest C_{llr}, and significantly improve performance in terms of C_{prim} with respect to linear models. Concerning short utterances, Log-Reg + QM_4 reduces both C_{llr} and EER with respect to standard Log-Reg, but slightly degrades the actual primary cost, although, given the small number of evaluation trials, this may not be statistically significant. On the other hand, VΓ-Var + Dur models outperform both Log-Reg + QM_4 and VΓ-Var both in terms of C_{llr} and C_{prim}, while providing similar EER.

Comparing the results of PLDA and VΓ-Var + Dur with FPD-PLDA and VΓ-Var, we can observe that the two models achieve similar performance. This confirms again that our duration model is indeed able to incorporate embedding uncertainty directly at calibration level, achieving similar effects as the combination of uncertainty models and PLDA back-ends, but without requiring an explicit model for embedding uncertainty. Since DNNs usually are not able to provide uncertainty measures, the VΓ-Var approach can be employed as a proxy for modeling the effects of uncertainty propagation directly at score level. As expected, duration-aware models do not further improve the FPD-PLDA results. Finally, results on long utterances show that duration modeling does not provide improvements in this scenario. Again, this was expected, since for longer utterances the miscalibration effects due to duration variability fade. The results are consistent with those of FPD-PLDA models for long segments [28].

Table V shows the results of PLDA with length-normalized embeddings and VΓ-Var calibration, and PLDA with non-normalized embeddings paired with VΓ-Var + Norm. The first four rows of the table refer to the original (long) segments. Without length normalization we observe a significant degradation of performance, with Log-Reg models providing significantly worse C_{prim} than VΓ-Var. Once we incorporate the embedding norm at calibration level, however, the gap is significantly reduced. We observe only a small degradation in terms of C_{llr} and EER with respect to length-normalized PLDA, although the degradation in terms of C_{prim} is still relevant. The second part of the table compares the same approaches over short segments. Again, length normalization is essential for this dataset, and also in this case VΓ-Var + Norm significantly reduces the gap between non-normalized and normalized embeddings. Finally, Rows 9 to 12 report the results for duration-aware models. Results are similar to those obtained for long segments. In contrast with SRE 2019 tests, in this case we observe an improvement when we combine both duration and embedding norms, although length normalization paired with VΓ-Var + Dur models seems to be more effective than the combined VΓ-Var + Norm + Dur approach.

E. SRE 2012

The last set of experiments was conducted on SRE 2012 evaluation data. The SRE 2012 enrollment set consists of multiple repetitions per speaker, some being the same utterance recorded through different microphones. Since these utterances are not independent, it's difficult to define an effective utterance duration for the enrollment side. However, as we showed in [29], we can assume that enrollment embeddings have been extracted from long utterances, i.e. we can ignore the uncertainty due to the enrollment duration. Table VI reports the results of different calibration approaches. Duration-aware models have been trained with a fixed enrollment duration, set to a large value. The results show that for this dataset Logistic Regression based models are effective and provide good calibration. The VΓ-Var models do not provide improvements in terms of C_{llr}, but improve calibration for low false-alarm regions, resulting in lower primary costs. Concerning duration-aware models, we observe that the proposed approach achieves again similar C_{llr} as Log-Reg + QM_4, while slightly decreasing the actual primary costs.

TABLE IV: Results on SRE 2010 with an i-vector frontend.

	Short segments						Original (long) segments		
	PLDA			FPD-PLDA			PLDA		
	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER	C_{llr}	C_{prim}	EER
Min. costs	*0.288*	*0.756*	*8.1%*	*0.251*	*0.764*	*7.3%*	*0.094*	*0.420*	*2.5%*
Log-Reg	0.297	0.818	8.1%	0.260	0.809	7.3%	0.098	0.444	2.5%
Linear VΓ	0.298	0.835	8.1%	0.260	0.818	7.3%	0.099	0.460	2.5%
VΓ-Var (Gen)	0.292	0.791	8.1%	0.256	0.797	7.3%	0.097	0.430	2.5%
VΓ-Var (Disc)	0.292	0.784	8.1%	0.257	0.774	7.3%	0.097	0.425	2.5%
Log+Reg + QM_4	0.261	0.827	7.0%	0.260	0.832	7.2%	0.098	0.451	2.5%
VΓ-Var + Dur (Gen)	0.259	0.755	7.0%	0.256	0.798	7.3%	0.097	0.430	2.5%
VΓ-Var + Dur (Disc)	0.253	0.787	7.0%	0.257	0.774	7.3%	0.097	0.425	2.5%

TABLE V: Results on SRE 2010 with an i-vector frontend.

PLDA L2-norm	Calibration Model	C_{llr}	C_{prim}	EER
Original (long) segments				
✓	VΓ-Var (Gen.)	0.097	0.430	2.5%
✗	LogReg	0.173	0.612	4.3%
✗	VΓ-Var (Gen.)	0.167	0.491	4.3%
✗	VΓ-Var + Norm (Gen.)	0.102	0.501	2.6%
Short segments				
✓	VΓ-Var (Gen.)	0.292	0.791	8.1 %
✗	LogReg	0.427	0.971	12.7%
✗	VΓ-Var (Gen.)	0.415	0.914	12.7%
✗	VΓ-Var + Norm (Gen.)	0.305	0.789	8.7%
✓	VΓ-Var + Dur (Gen.)	0.259	0.755	7.0%
✗	LogReg + QM_4	0.345	0.932	9.5%
✗	VΓ-Var + Dur (Gen.)	0.333	0.874	9.4%
✗	VΓ-Var + Norm + Dur (Gen.)	0.282	0.790	8.0%

TABLE VI: Results on SRE 2012 with an e-vector frontend.

	PLDA		PSVM	
	C_{llr}	C_{prim}	C_{llr}	C_{prim}
Min. costs	*0.062*	*0.211*	*0.057*	*0.210*
Log-Reg	0.065	0.244	0.061	0.217
Linear VΓ	0.066	0.269	0.061	0.224
VΓ-Var (Gen)	0.066	0.241	0.062	0.213
VΓ-Var (Disc)	0.065	0.224	0.061	0.212
Log+Reg + QM_4	0.054	0.211	0.051	0.209
VΓ-Var + Dur (Gen)	0.056	0.230	0.054	0.204
VΓ-Var + Dur (Disc)	0.055	0.186	0.053	0.199

VII. Conclusions

We have presented a generative model able to incorporate utterance-dependent miscalibration sources in terms of "effective", utterance-dependent within-class variance. This allows us, for example, to explicitly model utterance duration at calibration level. The resulting 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 of the VΓ-Var approach for semi-supervised scenarios. Furthermore, our approach provides strong interpretations for the calibration parameters, and a novel interpretation of the role of embedding norm in representing utterance-level uncertainty. Although the model assumes common between-class variability for all speakers, the approach can be easily extended to deal with sub-groups of speakers with different between class variability (e.g. male and female gender). Future work will also investigate the effects of employing variable between-speaker variability factors, and other methods to measure utterance-dependent variability, including, for example, those provided by score normalization statistics [52], with the goal of providing a unified framework for score normalization and calibration.

References

[1] N. Brummer *et 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.

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

[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*. PhD thesis, Autonomous University of Madrid, 2007.

[6] M. I. Mandasari, M. Günther, 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*, pp. 1619–1623, 2013.

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

[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*, pp. 14–18, 2014.

[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*, pp. 6121–6125, 05 2019.

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

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

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

[19] L. Ferrer, M. McLaren, and N. Brümmer, "A speaker verification backend with robust performance across conditions," *Computer Speech & Language*, vol. 71, p. 101258, 2022.

[20] 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*, pp. 5329–5333, 2018.

[21] S. Cumani and S. Sarni, "A generative model for duration-dependent score calibration," in *Proc. Interspeech 2021*, pp. 4598–4602, 2021.

[22] 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.

[23] 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.

[24] 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.

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

[26] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey 2010*, pp. 194–201, 2010.

[27] 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.

[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*, pp. 7644–7648, 2013.

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

[31] Z. Tavares, X. Zhang, E. Minaysan, J. Burroni, R. Ranganath, and A. Solar-Lezama, "The random conditional distribution for higher-order probabilistic inference," *CoRR*, vol. abs/1903.10556, 2019.

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

[33] 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.

[34] 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*, pp. 4832–4835, 2011.

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

[36] 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.

[37] N. Brummer, A. Silnova, L. Burget, and T. Stafylakis, "Gaussian meta-embeddings for efficient scoring of a heavy-tailed plda model," in *Proceedings of Odyssey 2018*, pp. 349–356, 06 2018.

[38] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[39] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015. arXiv:1510.08484v1.

[40] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, 2017.

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

[42] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *INTERSPEECH*, 2018.

[43] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations," *Comput. Speech Lang.*, vol. 60, 2020.

[44] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *INTERSPEECH*, 2020.

[45] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019.

[46] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware Deep Neural Networks," in *Proceedings of ICASSP 2014*, pp. 1695–1699, 2014.

[47] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep Neural Networks for extracting Baum-Welch statistics for speaker recognition," in *Proceedings of Odyssey 2014*, pp. 293–298, 2014.

[48] S. Cumani, P. Laface, and F. Kulsoom, "Speaker recognition by means of acoustic and phonetically informed GMMs," in *Proceedings of Interspeech 2015*, pp. 200–204, 2015.

[49] N. Brümmer, *Measuring, refining and calibrating speaker and language information extracted from speech*. PhD thesis, Stellenbosch University, South Africa, 2010.

[50] 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.

[51] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2002.

[52] S. Cumani and S. Sarni, "Impostor score statistics as quality measures for the calibration of speaker verification systems," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, pp. 25–32, 2022.