POLITECNICO DI TORINO
Repository ISTITUZIONALE

On the Distribution of Speaker Verification Scores: Generative Models for Unsupervised Calibration

Original
On the Distribution of Speaker Verification Scores: Generative Models for Unsupervised Calibration / Cumani, S.. - In: IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. - ISSN 2329-9290. - ELETTRONICO. - 29:(2021), pp. 547-562. [10.1109/TASLP.2020.3040103]

Availability:
This version is available at: 11583/2872304 since: 2021-02-23T17:37:45Z

Publisher:
IEEE

Published
DOI:10.1109/TASLP.2020.3040103

Terms of use:

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

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

15 May 2024

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

1

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

On the distribution of speaker verification scores: generative models for unsupervised calibration

Sandro Cumani

***Abstract*—Speaker verification systems whose outputs can be interpreted as log-likelihood ratios (LLR) allow for cost-effective decisions by comparing the system outputs to application-defined thresholds depending only on prior information. Classifiers often produce uncalibrated scores, and require additional processing to produce well-calibrated LLRs. Recently, generative score calibration models have been proposed, which achieve calibration performance close to that of state-of-the-art discriminative techniques for supervised scenarios, while also allowing for unsupervised training. The effectiveness of these methods, however, strongly depends on their capabilities to correctly model the target and non-target score distributions.**

In this work we propose theoretically grounded and accurate models for characterizing the distribution of scores of speaker verification systems. Our approach is based on tied Generalized Hyperbolic distributions and overcomes many limitations of Gaussian models. Experimental results on different NIST benchmarks, using different utterance representation front-ends and different back-end classifiers, show that our method is effective not only in supervised scenarios, but also in unsupervised tasks characterized by very low proportion of target trials.

***Index Terms*—Score calibration, log-likelihood ratio, unsupervised training, generalized hyperbolic distribution, variance-gamma distribution**

I. Introduction

Well-calibrated speaker verification systems output scores that can be interpreted as the log-likelihood ratio (LLR) between the same-speaker (target trial) and different-speaker (non-target trial) hypotheses. Given an application, hard decisions can be taken comparing the score with a suitable threshold. If a score is a LLR, the optimal threshold depends only on the prior probability of the two hypotheses, and the costs of the false acceptance and false rejection errors. In practice, most speaker verification systems are not able to directly produce well-calibrated scores. This may depend on several reasons, for example on the intrinsic nature of the classifier, on mismatches between the training and evaluation populations, or on imprecise model assumptions. For these reasons, score calibration techniques are employed to transform the scores produced by a recognizer so that they approximate well-calibrated LLRs.

The standard approach for score calibration is based on discriminative prior-weighted Logistic Regression (Log-Reg) [1], [2], routinely employed as a calibration tool for different tasks [3]–[6]. Since this approach is supervised, it requires a labeled dataset that closely matches the testing conditions. Recently, alternative methods based on generative models have been proposed [7]–[10]. These models not only provide more insights on the behavior of well-calibrated LLRs, but can be easily extended to handle missing labels. The work [7] shows the constraints that theoretical distributions of well calibrated LLRs should satisfy. The authors then propose the Constrained Maximum Likelihood Gaussian (CMLG) linear calibration model. It assumes that observed scores are linearly transformed samples of well-calibrated, Gaussian distributed, random variables satisfying the LLR constraints. An unsupervised extension of CMLG was proposed in [11].

CMLG can achieve good calibration, but its effectiveness depends on the accuracy of its Gaussian assumptions. Empirical score distributions often exhibit skewed, asymmetric and heavier-than-Gaussian tail behavior. In these cases, CMLG cannot properly model the score distributions, thus producing significant calibration loss. The behavior of empirical score distributions should not be surprising: as we show in Section III, even under the assumptions of a simplified Probabilistic Linear Discriminant Analysis (PLDA) [12], [13] model, well-calibrated LLRs are not Gaussian, but rather Variance-Gamma (VΓ) distributed. This issue is more relevant for unsupervised tasks. Since non-target scores are usually the vast majority, target scores can be easily confused as scores generated from the non-target distribution tails. Accurate modeling of the score distribution tails is thus very important. Alternative non-Gaussian models have been investigated in [8], using Normal Inverse Gaussian (NIG) densities. Contrary to CMLG, this approach does not assume a specific calibration model, but estimates a probabilistic model in score space. Well-calibrated scores are computed as the log-likelihood ratio between the hypotheses that a score was generated by the target or by the non-target distribution, respectively. NIG densities allow for better characterization of the score distributions, and for better accuracy with respect to CMLG for supervised tasks. However, as shown in [10], and confirmed in Section VI, they can lead to poor results for unsupervised scenarios, without appropriate constraints on the calibration model. In [9], [10] we proposed the Constrained Maximum Likelihood NIG (CMLNIG in [10], C-NIG in the following) approach, which combines the benefits of NIG distributions with linear calibration assumptions.

In this work we provide a theoretical analysis of the distributions of PLDA-based verification scores, showing that Variance-Gamma (VΓ) densities are good candidates for modeling score distributions, and we propose the Constrained VΓ (C-VΓ) as alternative to CMLG and C-NIG. Since both VΓ and NIG belong to the family of Generalized Hyperbolic (GH) distributions [14], we consider a more general framework, introducing the Constrained GH (C-GH) method that allows implementing both C-NIG and C-VΓ as special cases. The model assumes that target and non-target scores are obtained as

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

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

an affine transformation of random samples of GH-distributed random variables, with parameters appropriately tied to satisfy the LLR constraints. We validate our approach on different datasets (SRE 2019 [15], SRE 2012 [16] and SRE 2010 [17]), with different front-ends (i-vectors [18], e-vectors [19], x-vectors [20]), different classifiers (PLDA [12], [13], Non-Linear PLDA [21], [22] and Pairwise Support Vector Machines [23], [24]), and different proportions of target to non-target training samples. The experimental results show that the C-VΓ approach not only outperforms both the CMLG and C-NIG approaches, but it achieves close to optimal calibration even in scenarios characterized by very low target proportions.

The rest of the paper is organized as follows. Section II recalls the CMLG model [7] and its unsupervised extension [11]. Section III investigates the distribution of well-calibrated scores computed by a PLDA model, showing that Variance-Gamma distributions are good candidates for generative calibration models. Section IV presents the supervised and unsupervised C-VΓ and C-GH models and their relationship with C-NIG. Section V describes the model parameter estimation procedure. The experimental results are reported in Section VI for supervised and unsupervised tasks, and conclusions are drawn in Section VII.

II. Gaussian models for linear score calibration

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

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

where e represents a trial (e.g., a pair of i-vectors [18] 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. Since speaker verification back-ends often do not output well calibrated scores, calibration techniques are used to estimate a transformation f that approximates the mapping of an uncalibrated score s to a well-calibrated score $x = f_{cal}(s)$.

A. Generative score models

Generative approaches estimate f_{cal} through a statistical model $\mathcal{M}'$ that describes the distribution of the observed scores. The model interprets 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 computes the log-likelihood ratio for the score under the two hypotheses:

$$x' = f_{cal}(s) = \log \frac{f_{S|\mathfrak{S}}(s)}{f_{S|\mathfrak{D}}(s)} \ . \tag{2}$$

B. The LLR constraint

In [7], [25] the authors show that well-calibrated scores satisfy the "LLR of the LLR is the LLR" property, i.e., for a well-calibrated score $x = LLR(e)$,

$$f_{cal}(x) = \log \frac{f_{X|\mathfrak{S}}(x)}{f_{X|\mathfrak{D}}(x)} = x = \log \frac{P(e|\mathfrak{S}, \mathcal{M})}{P(e|\mathfrak{D}, \mathcal{M})} \ , \tag{3}$$

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

where X denotes the R.V. responsible for the generation of score x. The LLR property constrains the admissible densities for target and non-target score distributions of well-calibrated scores. Indeed, from (3) it follows that:

$$f_{X|\mathfrak{S}}(s) = e^x f_{X|\mathfrak{D}}(s) \ . \tag{4}$$

This constraint can be expressed in terms of Moment Generating Functions (MGF):

$$M_{X|\mathfrak{D}}(t) = \mathbb{E}_{X|\mathfrak{D}}\left[e^{tX}\right] \ , \qquad M_{X|\mathfrak{S}}(t) = \mathbb{E}_{X|\mathfrak{S}}\left[e^{tX}\right] \ , \tag{5}$$

Assuming[2] that $M_{X|\mathfrak{D}}$ is well defined in an open set $\mathcal{O}_{\mathfrak{D}} \supset [0, 1]$ and $M_{X|\mathfrak{S}}$ is well defined in an open set $\mathcal{O}_{\mathfrak{S}} \supset [-1, 0]$, the LLR constraint can be expressed as [10]:

$$M_{X|\mathfrak{S}}(t) = \int e^{tx} f_{X|\mathfrak{S}}(x)dx = \int e^{(t+1)x} f_{X|\mathfrak{D}}(x)dx = M_{X|\mathfrak{D}}(t+1). \tag{6}$$

The LLR constraint in (4) requires that $f_{X|\mathfrak{S}}$ and $f_{X|\mathfrak{D}}$ are proper densities, i.e. $\int f_{X|\mathfrak{S}}(x)dx = \int f_{X|\mathfrak{D}}(x)dx = 1$. In terms of MGFs, this requirement corresponds to $M_{X|\mathfrak{D}}(1) = M_{X|\mathfrak{S}}(0) = 1$ and $M_{X|\mathfrak{S}}(-1) = M_{X|\mathfrak{D}}(0) = 1$. Equation (6) will be used to derive the expression for the target and the non-target densities in Section IV.

C. CMLG

The CMLG approach [7] models uncalibrated scores S as samples of well-calibrated R.V.s $X|\mathfrak{S}$ and $X|\mathfrak{D}$, transformed by an affine function $s = f_{cal}^{-1}(x) = \frac{x-b}{a}$, leading to the calibration transformation $f_{cal}(s) = as + b$. As the name suggests, the conditional distribution $X|\mathfrak{D}$ for well-calibrated non-target scores is assumed to be Gaussian. Imposing the LLR constraint (4) implies that also the target distribution is Gaussian, and that the two densities are controlled by a single free parameter μ:

$$f_{X|\mathfrak{S}}(x) = \mathcal{N}(x|\mu, 2\mu) \ , \quad f_{X|\mathfrak{D}}(x) = \mathcal{N}(x|-\mu, 2\mu) \tag{7}$$

The R.V.s that describe the observed scores are given by $S|\mathfrak{S} = f_{cal}^{-1}(X|\mathfrak{S})$ and $S|\mathfrak{D} = f_{cal}^{-1}(X|\mathfrak{D})$, distributed as:

$$f_{S|\mathfrak{S}}(s) = \mathcal{N}(s|m_{\mathfrak{S}}, v) \ , \quad f_{S|\mathfrak{D}}(s) = \mathcal{N}(s|m_{\mathfrak{D}}, v) \tag{8}$$

where $m_{\mathfrak{S}}, m_{\mathfrak{D}}$ and v are related to the model and calibration parameters μ, a and b by:

$$m_{\mathfrak{S}} = \frac{\mu}{a} - \frac{b}{a} \ , \quad m_{\mathfrak{D}} = -\frac{\mu}{a} - \frac{b}{a} \ , \quad v = \frac{2\mu}{a^2} \ . \tag{9}$$

CMLG estimates the parameters $m_{\mathfrak{S}}, m_{\mathfrak{D}}$ and v that maximize a weighted log-likelihood criterion for the observed scores:

$$\mathscr{L} = \frac{\zeta}{n_{\mathfrak{S}}} \sum_{i \in \mathcal{I}_{\mathfrak{S}}} \log \mathcal{N}(s_i|m_{\mathfrak{S}}, v) + \frac{1-\zeta}{n_{\mathfrak{D}}} \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \log \mathcal{N}(s_i|m_{\mathfrak{D}}, v) \tag{10}$$

where $\mathcal{I}_{\mathfrak{S}}$ and $\mathcal{I}_{\mathfrak{D}}$ denote the sets of indices, and $n_{\mathfrak{S}}$ and $n_{\mathfrak{D}}$ the corresponding cardinalities, of target and non-target scores, respectively. ζ is a weight that allows balancing the contribution of the scores of the different classes to the objective function. Given $m_{\mathfrak{S}}, m_{\mathfrak{D}}$ and v, from (9) one can get the parameters of the calibration function $f_{cal}(s) = as + b$.

[2]These assumptions guarantee that both MGFs are well-defined around 0 and thus the uniqueness of the corresponding probability density function.

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

D. Unsupervised CMLG

An unsupervised extension of CMLG was presented in [11]. The authors introduce a latent R.V. that represents the (unknown) label. The conditional distributions given the labels are derived from CMLG. Well-calibrated scores are therefore samples of R.V. X with density:

$$f_X(x) = \pi\mathcal{N}(x|\mu, 2\mu) + (1-\pi)\mathcal{N}(x|-\mu, 2\mu) \ . \tag{11}$$

where π is the prior probability for the target class. The observed scores are again interpreted as samples of the affine-transformed R.V. $S = f_{cal}^{-1}(X)$:

$$f_S(s) = \pi\mathcal{N}(s|m_{\mathfrak{S}}, v) + (1-\pi)\mathcal{N}(s|m_{\mathfrak{D}}, v) \ . \tag{12}$$

where $m_{\mathfrak{S}}$, $m_{\mathfrak{D}}$ and v have the same meaning as in (8). The model parameters, which include the additional weight π, can be estimated by maximizing the log-likelihood:

$$\mathscr{L} = \sum_{i\in\mathcal{I}} \log f_S(s_i) \ , \tag{13}$$

where $\mathcal{I}$ denotes the set of all score indices $\mathcal{I} = \mathcal{I}_{\mathfrak{S}} \cup \mathcal{I}_{\mathfrak{D}}$. Since labels are not available, the objective function does not allow for different weights for target and non-target samples.

The CMLG approach allows obtaining good results whenever the target and non-target score distributions have similar, approximately Gaussian, shape. However, empirical score distributions are usually asymmetric, skewed and present heavier-than-Gaussian tail behavior. For example, Figure 1 plots the histogram of target and non-target scores for the SRE 2019 Progress set using an x-vector front-end and a PLDA classifier. The non-target score distribution presents a heavier left tail, and the shapes of the two histograms are significantly different. In these cases, CMLG is less effective, especially for unsupervised scenarios with highly unbalanced classes.

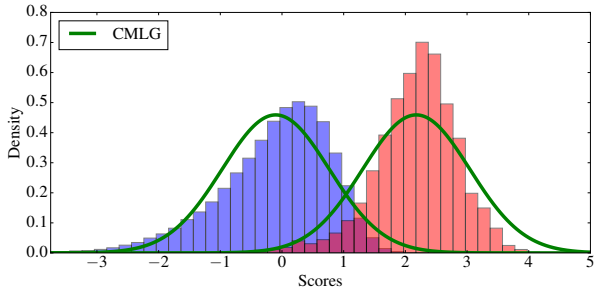


Fig. 1: Target (red) and non-target (blue) scores histogram for the SRE 2019 Progress set using an x-vector front-end with a PLDA classifier, and CMLG estimates of the target and non-target densities.

III. The distribution of Log-Likelihood Ratios of PLDA-distributed Speaker Vectors

To better capture the characteristics of the target and non-target scores in this section we analyze the distribution of well-calibrated scores generated by a PLDA back-end. We consider the simplified two-covariance version of the PLDA model[3]

$$\mathbf{\Phi} = \overline{\mathbf{Y}} + \overline{\mathbf{E}} \ , \tag{14}$$

where $\mathbf{\Phi}$ is the R.V. responsible for generating an observed speaker vector (e.g. i-vector or speaker embedding), $\overline{\mathbf{Y}}$ is the R.V. representing the speaker identity and $\overline{\mathbf{E}}$ represents residual noise. The prior distributions of $\overline{\mathbf{Y}}$ and $\overline{\mathbf{E}}$ are:

$$\overline{\mathbf{Y}} \sim \mathcal{N}(\mathbf{m}, \mathbf{B}) \ , \qquad \overline{\mathbf{E}} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}) \ . \tag{15}$$

Without loss of generality[4], we also assume that $\mathbf{m} = 0$, and that both $\mathbf{B}$ and $\mathbf{W}$ are diagonal. We consider pairs of speaker vectors $(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ that are generated by model (14). Under the same-speaker hypothesis, the i-vectors are sampled from

$$\begin{bmatrix}\mathbf{\Phi}_1 \\ \mathbf{\Phi}_2\end{bmatrix} |\mathfrak{S} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{B}+\mathbf{W} & \mathbf{B} \\ \mathbf{B} & \mathbf{B}+\mathbf{W}\end{bmatrix}\right) \ . \tag{16}$$

Under the different-speaker hypothesis, the pair distribution is

$$\begin{bmatrix}\mathbf{\Phi}_1 \\ \mathbf{\Phi}_2\end{bmatrix} |\mathfrak{D} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{B}+\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}+\mathbf{W}\end{bmatrix}\right) \ . \tag{17}$$

The two speaker vectors are independent under the different speaker hypothesis, whereas they are correlated under the same speaker hypothesis. It is therefore useful to recast the model in terms of R.V.s that are independent under both assumptions. We consider the R.V.s:

$$\mathbf{Z}_+ = \frac{(\mathbf{\Phi}_1 + \mathbf{\Phi}_2)}{\sqrt{2}}, \quad \mathbf{Z}_- = \frac{(\mathbf{\Phi}_1 - \mathbf{\Phi}_2)}{\sqrt{2}}, \quad \mathbf{Z} = \begin{bmatrix}\mathbf{Z}_+ \\ \mathbf{Z}_-\end{bmatrix} . \tag{18}$$

The distributions of $\mathbf{Z}$ conditioned on the same and different speaker hypothesis are, respectively:

$$\mathbf{Z}|\mathfrak{S} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\mathfrak{S}}\right) \ , \qquad \mathbf{Z}|\mathfrak{D} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\mathfrak{D}}\right) \ , \tag{19}$$

with covariance matrices:

$$\mathbf{\Sigma}_{\mathfrak{S}} = \begin{bmatrix}2\mathbf{B}+\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}\end{bmatrix}, \quad \mathbf{\Sigma}_{\mathfrak{D}} = \begin{bmatrix}\mathbf{B}+\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}+\mathbf{W}\end{bmatrix} \tag{20}$$

$\mathbf{\Sigma}_{\mathfrak{S}}$ and $\mathbf{\Sigma}_{\mathfrak{D}}$ are diagonal, and $\mathbf{Z}_+$ and $\mathbf{Z}_-$ are independent under both hypotheses. The LLR for a pair of speaker vectors $(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ can be written in terms of $\mathbf{z} = \begin{bmatrix}\mathbf{z}_+^T & \mathbf{z}_-^T\end{bmatrix}^T$, with

$$\mathbf{z}_+ = \frac{1}{\sqrt{2}}\left(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2\right) \ , \quad \mathbf{z}_- = \frac{1}{\sqrt{2}}\left(\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\right) \ , \tag{21}$$

as:

$$LLR(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \ell(\mathbf{z}) = \log \frac{f_{\mathbf{Z}|\mathfrak{S}}(\mathbf{z})}{f_{\mathbf{Z}|\mathfrak{D}}(\mathbf{z})} \ . \tag{22}$$

where $f_{\mathbf{Z}|\mathfrak{S}}$ and $f_{\mathbf{Z}|\mathfrak{D}}$ are the densities of $\mathbf{Z}|\mathfrak{S}$ and $\mathbf{Z}|\mathfrak{D}$. Replacing the densities with their expressions, we obtain that the LLR $\ell(\mathbf{z})$ is a quadratic function of $\mathbf{z}$:

$$\ell(\mathbf{z}) = -\frac{1}{2}\log\left|\mathbf{\Sigma}_{\mathfrak{S}}\mathbf{\Sigma}_{\mathfrak{D}}^{-1}\right| - \frac{1}{2}\mathbf{z}^T\left(\mathbf{\Sigma}_{\mathfrak{S}}^{-1} - \mathbf{\Sigma}_{\mathfrak{D}}^{-1}\right)\mathbf{z} \ . \tag{23}$$

[3]Our derivations can be easily extended to subspace-constrained models.

[4]These assumptions can be recovered through an affine transformation of the data, $f(\mathbf{\Phi}) = \mathbf{A}(\mathbf{\Phi} - \mathbf{m})$, where $\mathbf{A} = \mathbf{U}\mathbf{W}^{-\frac{1}{2}}$, with $\mathbf{U}$ given by the SVD of $\mathbf{W}^{-\frac{1}{2}}\mathbf{B}\mathbf{W}^{-\frac{1}{2}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

Function $\ell(\mathbf{z})$ maps pairs of speaker vectors to scores. If speaker vectors are samples of R.V. $\mathbf{Z}$, then the LLRs $\ell(\mathbf{z})$ are themselves samples of R.V. $\mathcal{L} = \ell(\mathbf{Z})$. The target and non-target scores are thus samples of $\mathcal{L}|\mathfrak{S} = \ell(\mathbf{Z}|\mathfrak{S})$ and $\mathcal{L}|\mathfrak{D} = \ell(\mathbf{Z}|\mathfrak{D})$. To derive the distribution of $\mathcal{L}|\mathfrak{D}$ we rewrite $\mathbf{Z}|\mathfrak{D}$ as:

$$\mathbf{Z}|\mathfrak{D} = \mathbf{\Sigma}_{\mathfrak{D}}^{\frac{1}{2}}\mathbf{Y} \ , \qquad \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ , \tag{24}$$

so that the R.V. $\mathcal{L}|\mathfrak{D}$ can be expressed as

$$\begin{aligned}\mathcal{L}|\mathfrak{D} &= K_{\mathbf{\Sigma}} - \frac{1}{2}(\mathbf{Z}|\mathfrak{D})^T \left(\mathbf{\Sigma}_{\mathfrak{S}}^{-1} - \mathbf{\Sigma}_{\mathfrak{D}}^{-1}\right)(\mathbf{Z}|\mathfrak{D}) \\ &= K_{\mathbf{\Sigma}} - \frac{1}{2}\mathbf{Y}^T\mathbf{\Sigma}_{\mathfrak{D}}^{\frac{1}{2}}\left(\mathbf{\Sigma}_{\mathfrak{S}}^{-1} - \mathbf{\Sigma}_{\mathfrak{D}}^{-1}\right)\mathbf{\Sigma}_{\mathfrak{D}}^{\frac{1}{2}}\mathbf{Y} \\ &= K_{\mathbf{\Sigma}} - \frac{1}{2}\mathbf{Y}^T\left(\mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}} - \mathbf{I}\right)\mathbf{Y} \ , \end{aligned} \tag{25}$$

where we set

$$\mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}} = \mathbf{\Sigma}_{\mathfrak{D}}^{\frac{1}{2}}\mathbf{\Sigma}_{\mathfrak{S}}^{-1}\mathbf{\Sigma}_{\mathfrak{D}}^{\frac{1}{2}} \ , \qquad K_{\mathbf{\Sigma}} = -\frac{1}{2}\log\left|\mathbf{\Sigma}_{\mathfrak{S}}\mathbf{\Sigma}_{\mathfrak{D}}^{-1}\right| \ . \tag{26}$$

Following a similar approach, we rewrite $\mathbf{Z}|\mathfrak{S}$ as

$$\mathbf{Z}|\mathfrak{S} = \mathbf{\Sigma}_{\mathfrak{S}}^{\frac{1}{2}}\mathbf{Y} \ , \qquad \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ , \tag{27}$$

and $\mathcal{L}|\mathfrak{S}$ is thus given by:

$$\begin{aligned}\mathcal{L}|\mathfrak{S} &= K_{\mathbf{\Sigma}} - \frac{1}{2}(\mathbf{Z}|\mathfrak{S})^T \left(\mathbf{\Sigma}_{\mathfrak{S}}^{-1} - \mathbf{\Sigma}_{\mathfrak{D}}^{-1}\right)(\mathbf{Z}|\mathfrak{S}) \\ &= K_{\mathbf{\Sigma}} - \frac{1}{2}\mathbf{Y}^T\mathbf{\Sigma}_{\mathfrak{S}}^{\frac{1}{2}}\left(\mathbf{\Sigma}_{\mathfrak{S}}^{-1} - \mathbf{\Sigma}_{\mathfrak{D}}^{-1}\right)\mathbf{\Sigma}_{\mathfrak{S}}^{\frac{1}{2}}\mathbf{Y} \\ &= K_{\mathbf{\Sigma}} - \frac{1}{2}\mathbf{Y}^T\left(\mathbf{I} - \mathbf{\Sigma}_{\mathfrak{S}/\mathfrak{D}}\right)\mathbf{Y} \end{aligned} \tag{28}$$

with

$$\mathbf{\Sigma}_{\mathfrak{S}/\mathfrak{D}} = \mathbf{\Sigma}_{\mathfrak{S}}^{\frac{1}{2}}\mathbf{\Sigma}_{\mathfrak{D}}^{-1}\mathbf{\Sigma}_{\mathfrak{S}}^{\frac{1}{2}} \ . \tag{29}$$

Recalling that $\mathbf{B}$ and $\mathbf{W}$ are diagonal, it follows that also $\mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}}$ and $\mathbf{\Sigma}_{\mathfrak{S}/\mathfrak{D}}$ are diagonal, with $\mathbf{\Sigma}_{\mathfrak{S}/\mathfrak{D}} = \mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}}^{-1}$, and

$$\mathrm{Diag}(\mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}}) = \left[\left[\frac{b_1+w_1}{2b_1+w_1} \cdots \frac{b_M+w_M}{2b_M+w_M}\right]\left[\frac{b_1+w_1}{w_1} \cdots \frac{b_M+w_M}{w_M}\right]\right] \tag{30}$$

where b_i and w_i are the i-th elements of the diagonal of $\mathbf{B}$ and $\mathbf{W}$, respectively, and M is the size of the feature space (i.e. the size of the speaker vectors $\boldsymbol{\phi}$).

Matrices $\mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}}$ and $\mathbf{\Sigma}_{\mathfrak{S}/\mathfrak{D}}$ depend on the between and within class variability only through the between-over-within variance ratios $\rho_i = \frac{b_i}{w_i}$:

$$\frac{b_i + w_i}{2b_i + w_i} = \frac{\rho_i + 1}{2\rho_i + 1} \ , \qquad \frac{b_i + w_i}{w_i} = \rho_i + 1 \ . \tag{31}$$

The matrices $\mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}} - \mathbf{I}$ and $\mathbf{I} - \mathbf{\Sigma}_{\mathfrak{S}/\mathfrak{D}}$ are thus diagonal, with:

$$\mathrm{Diag}(\mathbf{\Sigma}_{\mathfrak{D}/\mathfrak{S}} - \mathbf{I}) = \left[\left[-\frac{\rho_1}{2\rho_1+1} \cdots -\frac{\rho_M}{2\rho_M+1}\right], \left[\rho_1 \ldots \rho_M\right]\right] \tag{32}$$

and

$$\mathrm{Diag}(\mathbf{I} - \mathbf{\Sigma}_{\mathfrak{S}/\mathfrak{D}}) = \left[\left[-\frac{\rho_1}{\rho_1+1} \cdots -\frac{\rho_M}{\rho_M+1}\right]\left[\frac{\rho_1}{\rho_1+1} \cdots \frac{\rho_M}{\rho_M+1}\right]\right] \tag{33}$$

Rewriting the quadratic forms in (25) and (28) in terms of the components of $\mathbf{Y}$ we get:

$$\mathcal{L}|\mathfrak{D} = K_{\mathbf{\Sigma}} + \frac{1}{2}\left[\sum_{i=1}^{M}\frac{\rho_i}{2\rho_i+1}Y_{i,+}^2 - \sum_{i=1}^{M}\rho_i Y_{i,-}^2\right] \tag{34}$$

$$\mathcal{L}|\mathfrak{S} = K_{\mathbf{\Sigma}} + \frac{1}{2}\left[\sum_{i=1}^{M}\frac{\rho_i}{\rho_i+1}Y_{i,+}^2 - \sum_{i=1}^{M}\frac{\rho_i}{\rho_i+1}Y_{i,-}^2\right] \tag{35}$$

where $Y_{i,+}$ and $Y_{i,-}$ are 1-dimensional, independent, standard normal distributed R.V.s, corresponding to the components of the first and second half of $\mathbf{Y}$, respectively. Since each $Y_{i,\cdot}$ is standard normal distributed, we have that:

$$Y_{i,\cdot} \sim \mathcal{N}(0,1) \implies aY_{i,\cdot}^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2a}\right), \ a > 0. \tag{36}$$

where $\Gamma(a, b)$ denotes a Gamma distribution with shape a and rate b. $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ can be rewritten as

$$\mathcal{L}|\mathfrak{D} \sim K_{\mathbf{\Sigma}} + \sum_{i=1}^{M}G_{i,+}^{\mathfrak{D}} - \sum_{i=1}^{M}G_{i,-}^{\mathfrak{D}} \tag{37}$$

$$\mathcal{L}|\mathfrak{S} \sim K_{\mathbf{\Sigma}} + \sum_{i=1}^{M}G_{i,+}^{\mathfrak{S}} - \sum_{i=1}^{M}G_{i,-}^{\mathfrak{S}} \tag{38}$$

where $G_{\cdot,\cdot}$ are independent, Gamma-distributed R.V.s:

$$G_{i,+}^{\mathfrak{D}} \sim \Gamma\left(\frac{1}{2}, \frac{2\rho_i+1}{\rho_i}\right) \ , \qquad G_{i,-}^{\mathfrak{D}} \sim \Gamma\left(\frac{1}{2}, \frac{1}{\rho_i}\right) \tag{39}$$

$$G_{i,+}^{\mathfrak{S}} \sim \Gamma\left(\frac{1}{2}, \frac{\rho_i+1}{\rho_i}\right) \ , \qquad G_{i,-}^{\mathfrak{S}} \sim \Gamma\left(\frac{1}{2}, \frac{\rho_i+1}{\rho_i}\right) \tag{40}$$

The term $K_{\mathbf{\Sigma}}$ can also be expressed as a sum of M terms:

$$K_{\mathbf{\Sigma}} = \sum_{i=1}^{M}\mu_i \ , \qquad \mu_i = \log\left(\frac{(\rho_i+1)^2}{2\rho_i+1}\right)^{\frac{1}{2}} \ . \tag{41}$$

Replacing $K_{\mathbf{\Sigma}}$ in (37) and (38) we can write $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ as

$$\mathcal{L}|\mathfrak{D} = \sum_{i=1}^{M}\mathcal{L}_i|\mathfrak{D} \ , \qquad \mathcal{L}|\mathfrak{S} = \sum_{i=1}^{M}\mathcal{L}_i|\mathfrak{S} \tag{42}$$

where $\mathcal{L}_i|\mathfrak{D}$ and $\mathcal{L}_i|\mathfrak{S}$ are independent R.V.s:

$$\mathcal{L}_i|\mathfrak{D} = \mu_i + G_{i,+}^{\mathfrak{D}} - G_{i,-}^{\mathfrak{D}} \ , \quad \mathcal{L}_i|\mathfrak{S} = \mu_i + G_{i,+}^{\mathfrak{S}} - G_{i,-}^{\mathfrak{S}} \ . \tag{43}$$

The distributions of target and non-target LLRs of a M-dimensional PLDA model are given by the sum of M independent terms, each representing the contribution of a different speaker vector direction. Furthermore, these terms depend only on between-over-within class variance ratios. To characterize the LLR distributions we therefore focus our attention to the distribution of the different terms $\mathcal{L}_i|\mathfrak{D}$ and $\mathcal{L}_i|\mathfrak{S}$. Both variables correspond to a difference of two Gamma-distributed R.V.s with the same shape parameter. In [26] it was shown that the result is a R.V. with a Variance-Gamma distribution:

$$X \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta, \mu) \iff M_X(t) = e^{t\mu}\left(1 - \frac{2\beta}{\gamma^2}t - \frac{t^2}{\gamma^2}\right)^{-\lambda} \tag{44}$$

with $\gamma^2 = \alpha^2 - \beta^2 > 0$, $\alpha > 0$ and $t \in (-\alpha - \beta, \alpha - \beta)$.

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

This result derives from the properties of MGFs, and from the MGF of the Gamma distribution:

$$X \sim \Gamma(a,b) \iff M_X(t) = \left(1 - \frac{t}{b}\right)^{-a} \quad \text{for } t < b \ . \quad (45)$$

For the non-target components we have

$$\begin{aligned} M_{\mathcal{L}_i|\mathfrak{D}}(t) &= e^{t\mu_i} \cdot M_{G^{\mathfrak{D}}_{i,+}}(t) \cdot M_{G^{\mathfrak{D}}_{i,-}}(-t) \\ &= e^{t\mu_i}\left(1 + \frac{2\rho_i^2}{2\rho_i+1}t - \frac{\rho_i^2}{2\rho_i+1}t^2\right)^{-\frac{1}{2}} , \end{aligned} \quad (46)$$

for $t \in \left(-\frac{1}{\rho_i}, \frac{2\rho_i+1}{\rho_i}\right) \supset [0,1]$. The MGF of the target components is

$$M_{\mathcal{L}_i|\mathfrak{S}}(t) = e^{t\mu_i}\left(1 - \frac{\rho_i^2}{(\rho_i+1)^2}t^2\right)^{-\frac{1}{2}} \quad (47)$$

for $t \in \left(-\frac{\rho_i+1}{\rho_i}, \frac{\rho_i+1}{\rho_i}\right) \supset [-1,0]$. Both $\mathcal{L}_i|\mathfrak{D}$ and $\mathcal{L}_i|\mathfrak{S}$ are thus Variance-Gamma distributed:

$$\begin{aligned} \mathcal{L}_i|\mathfrak{D} &\sim \mathrm{V\Gamma}(\lambda_{i,\mathfrak{D}}, \beta_{i,\mathfrak{D}}, \alpha_{i,\mathfrak{D}}, \mu_{i,\mathfrak{D}}) \\ \mathcal{L}_i|\mathfrak{S} &\sim \mathrm{V\Gamma}(\lambda_{i,\mathfrak{S}}, \beta_{i,\mathfrak{S}}, \alpha_{i,\mathfrak{S}}, \mu_{i,\mathfrak{S}}) \ . \end{aligned} \quad (48)$$

The parameters can be inferred by inspection:

$\mathcal{L}_i|\mathfrak{D}$:

$$\begin{aligned} \mu_{i,\mathfrak{D}} &= \mu_i = \log\left(\frac{(\rho_i+1)^2}{2\rho_i+1}\right)^{\frac{1}{2}} \\ \gamma^2_{i,\mathfrak{D}} &= \frac{2\rho_i+1}{\rho_i^2} \\ \beta_{i,\mathfrak{D}} &= -1 \\ \alpha^2_{i,\mathfrak{D}} &= \gamma^2_{i,\mathfrak{D}} + 1 = \frac{(\rho_i+1)^2}{\rho_i^2} \\ \lambda_{i,\mathfrak{D}} &= \frac{1}{2} \end{aligned} \quad (49)$$

$\mathcal{L}_i|\mathfrak{S}$:

$$\begin{aligned} \mu_{i,\mathfrak{S}} &= \mu_i \\ \gamma^2_{i,\mathfrak{S}} &= \frac{(\rho_i+1)^2}{\rho_i^2} \\ \beta_{i,\mathfrak{S}} &= 0 \\ \alpha^2_{i,\mathfrak{S}} &= \gamma^2_{i,\mathfrak{S}} = \frac{(\rho_i+1)^2}{\rho_i^2} \\ \lambda_{i,\mathfrak{S}} &= \frac{1}{2} \end{aligned} \quad (50)$$

From (49) and (50) we can notice that, for a given direction i, the two distributions $\mathcal{L}_i|\mathfrak{D}$ and $\mathcal{L}_i|\mathfrak{S}$ share all parameters except $\beta_{i,\mathfrak{D}}$ and $\beta_{i,\mathfrak{S}}$. These parameters control the skewness of the distributions. In particular, the target distribution is not skewed, while the non-target distribution has negative skewness, corresponding to a longer left tail. Furthermore, the parameters only depend on the between-to-within class variance ratios ρ_i, and do not depend on the scale of the input space. Finally, we can observe that each pair $\mathcal{L}_i|\mathfrak{D}$ and $\mathcal{L}_i|\mathfrak{S}$ satisfies the LLR constraint (6), since there exists an open set $\mathcal{O}_i \supset [0,1]$ such that, for $t \in \mathcal{O}_i$,

$$M_{\mathcal{L}_i|\mathfrak{D}}(t) = M_{\mathcal{L}_i|\mathfrak{S}}(t-1) \ . \quad (51)$$

The R.V.s that model target and non-target scores $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ are sums of Variance-Gamma components with parameters specified by (49) and (50). In the general case the density of $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ cannot be written in closed form. Closed form expressions can be recovered if we assume isotropic between-over-within variance ratios:

$$\frac{b_i}{w_i} = \rho_i = \rho \ , \quad \forall i = 1, \dots, M \quad (52)$$

In this case, letting $\mu = \sum_{i=1}^{M} \mu_i = K_{\boldsymbol{\Sigma}}$, the MGF of $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ are

$$\begin{aligned} M_{\mathcal{L}|\mathfrak{D}} &= \prod_{i=1}^{M} M_{\mathcal{L}_i|\mathfrak{D}} = e^{t\mu}\left(1 + \frac{2\rho^2}{2\rho+1}t - \frac{\rho^2}{2\rho+1}t^2\right)^{-\frac{M}{2}} , \\ M_{\mathcal{L}|\mathfrak{S}} &= \prod_{i=1}^{M} M_{\mathcal{L}_i|\mathfrak{S}} = e^{t\mu}\left(1 - \frac{\rho^2}{(\rho+1)^2}t^2\right)^{-\frac{M}{2}} , \end{aligned} \quad (53)$$

corresponding to the Variance-Gamma distributions

$$\mathcal{L}|\mathfrak{D} \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta_{\mathfrak{D}}, \mu) \ , \quad \mathcal{L}|\mathfrak{S} \sim \mathrm{V\Gamma}(\lambda, \alpha, \beta_{\mathcal{S}}, \mu) \quad (54)$$

where

$$\begin{aligned} \beta_{\mathfrak{D}} &= -1 \ , & \alpha^2 &= \frac{(\rho+1)^2}{\rho^2} \ , \\ \beta_{\mathfrak{S}} &= 0 \ , & \lambda &= \frac{M}{2} \ . \end{aligned} \quad (55)$$

As expected, also $\mathcal{L}|\mathfrak{D}$ and $\mathcal{L}|\mathfrak{S}$ satisfy the LLR constraint. Let $\mathcal{O}$ be the smallest open set that includes all sets $\mathcal{O}_i$. The LLR constraint is satisfied since $[0,1] \subset \mathcal{O}$, and, for $t \in \mathcal{O}$,

$$M_{\mathcal{L}|\mathfrak{D}}(t) = \prod_i M_{\mathcal{L}_i|\mathfrak{D}}(t) = \prod_i M_{\mathcal{L}_i|\mathfrak{S}}(t-1) = M_{\mathcal{L}|\mathfrak{S}}(t-1) \quad (56)$$

We have so far shown that, even for well-calibrated PLDA systems, scores are not Gaussian distributed, but rather present skewness and semi-heavy tailed behavior. In presence of miscalibration sources the score distribution will, in general, present different shapes than those we investigated in this Section. However, the effectiveness of discriminative linear calibration approaches suggests that, except for location and scale, the shape of the score distributions does not change significantly with respect to well-calibrated scores in many practical cases. This motivates us to investigate the use of Variance-Gamma densities for modeling target and non-target distributions, following an approach similar to CMLG. The method we propose, Constrained ML Variance Gamma (C-VΓ), shares many aspects with our previously proposed approach based on Normal Inverse Gaussian (NIG) distributions [9], [10]. Indeed, both NIG and VΓ belong to the family of Generalized Hyperbolic (GH) distributions. Furthermore, we observed that estimated NIG densities provide good approximations of VΓ densities estimated from verification scores. This can explain the effectiveness of the Constrained Maximum Likelihood NIG (C-NIG) [10] method. As we show in the experimental section, however, the C-VΓ approach outperforms not only the CMLG approach, but also the C-NIG one. Additionally, given the effectiveness of both C-VΓ and C-NIG, we investigate the broader class of Generalized Hyperbolic distributions. This will allow us to unify the estimation procedures for both approaches.

IV. Constrained Generalized Hyperbolic models for Linear Calibration

The Generalized Hyperbolic distribution [14] belongs to the family of Gaussian mean-variance mixtures, i.e., distributions that are defined as:

$$X = \mu + \beta V + \sqrt{V}Y \quad (57)$$

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

where V and Y are independent R.V.s, Y is standard normal distributed and the density of the mixing variable V is $f_V(v)$.

A. Normal Inverse Gaussian model

Score calibration based on constrained Gaussian mean-variance mixtures was first analyzed in [9], where sufficient conditions were derived for target and non-target densities of well-calibrated R.V.s

$$\begin{aligned} X|\mathfrak{S} &= \mu_{\mathfrak{S}} + \beta_{\mathfrak{S}} V_{\mathfrak{S}} + \sqrt{V_{\mathfrak{S}}} Y \\ X|\mathfrak{D} &= \mu_{\mathfrak{D}} + \beta_{\mathfrak{D}} V_{\mathfrak{D}} + \sqrt{V_{\mathfrak{D}}} Y \end{aligned} \tag{58}$$

We proved that, as long as the mixing densities are equal, $f_{V_{\mathfrak{S}}} = f_{V_{\mathfrak{D}}}$, and the parameters are tied ($\mu_{\mathfrak{S}} = \mu_{\mathfrak{D}} = 0$, and $\beta_{\mathfrak{S}} = -\beta_{\mathfrak{D}} = \frac{1}{2}$), the distributions in (58) satisfy the LLR constraint. However, these conditions correspond to target distributions that are symmetric to non-target densities. In [10] we further showed that these conditions can be relaxed to fit a wider range of possible score distributions. In this Section we briefly recall the relaxed C-NIG model of [10].

The Normal Inverse Gaussian [27] is a 3-parameter distribution corresponding to a mean-variance mixture (57) where the mixing distribution is an Inverse Gaussian (IG). The NIG density is given by:

$$f_{\text{NIG}}(x|\alpha,\beta,\delta,\mu) = \frac{\alpha\delta K_1\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)}{\pi\sqrt{\delta^2+(x-\mu)^2}} e^{\delta\gamma+\beta(x-\mu)} , \tag{59}$$

where μ is a location parameter, $\delta > 0$ is a scaling parameter, β controls the skewness of the distribution, $\alpha > |\beta|$ controls the heaviness of the distribution tails, and $\gamma = \sqrt{\alpha^2-\beta^2}$. $K_\nu(x)$ is the modified Bessel function of the third kind of order ν. The C-NIG model assumes that target and non-target scores s_i are obtained as an affine transformation of well-calibrated scores x_i, $s_i = \frac{x_i - b}{a}$. The calibrated scores are sampled from NIG-distributed R.V.s:

$$\begin{aligned} X|\mathfrak{D} &\sim \text{NIG}\left(\alpha,\beta,\delta,\delta\left(\gamma_{\mathfrak{S}}-\gamma_{\mathfrak{D}}\right)\right) , \\ X|\mathfrak{S} &\sim \text{NIG}\left(\alpha,\beta+1,\delta,\delta\left(\gamma_{\mathfrak{S}}-\gamma_{\mathfrak{D}}\right)\right) , \end{aligned} \tag{60}$$

where $\alpha > \max(|\beta|, |\beta+1|)$, $\gamma_{\mathfrak{S}} = \sqrt{\alpha^2-(\beta+1)^2}$ and $\gamma_{\mathfrak{D}} = \sqrt{\alpha^2-\beta^2}$. The location parameter μ is tied to the other parameters to satisfy the LLR constraint. As in CMLG, the R.V.s that generates the observed scores are:

$$S|\mathfrak{S} = \frac{1}{a}(X|\mathfrak{S}) - \frac{b}{a} , \quad S|\mathfrak{D} = \frac{1}{a}(X|\mathfrak{D}) - \frac{b}{a} , \tag{61}$$

and they are also NIG-distributed, with parameters depending on the free parameters α, β, δ and the calibration parameters a, b. As we show in Section IV-D, the C-NIG model is a special case of the Constrained GH model.

B. Variance-Gamma model

As we have shown in Section III the Variance-Gamma distribution corresponds to the distribution of well-calibrated scores generated from a PLDA model with isotropic between-over-within class variance ratio. VΓ distributions are Gaussian mean-variance mixtures (57), where the mixing R.V. has a Gamma density. The result is a 4-parameter family:

$$f_{\text{V}\Gamma}(x|\lambda,\alpha,\beta,\mu) = \frac{\gamma^{2\lambda}|x-\mu|^{\lambda-\frac{1}{2}} K_{\lambda-\frac{1}{2}}(\alpha|x-\mu|)}{\sqrt{\pi}\Gamma(\lambda)(2\alpha)^{\lambda-\frac{1}{2}}} e^{\beta(x-\mu)} \tag{62}$$

where μ, β and $\alpha > |\beta|$ control the location, skewness and tail heaviness, respectively, and λ is a shape parameter. Following the CMLG and C-NIG approach, we assume that well-calibrated scores are generated according to the VΓ distributions:

$$X|\mathfrak{D} \sim \text{V}\Gamma(\lambda,\alpha,\beta_{\mathfrak{D}},\mu) , \quad X|\mathfrak{S} \sim \text{V}\Gamma(\lambda,\alpha,\beta_{\mathfrak{S}},\mu) . \tag{63}$$

According to model (54), μ and α both depend on the single parameter ρ, and are therefore tied, whereas λ depends only on the speaker vector dimensionality, $\lambda = \frac{M}{2}$, and $\beta_{\mathfrak{D}}$ and $\beta_{\mathfrak{S}}$ should be fixed to -1 and 0, respectively. Letting $\gamma_{\mathfrak{S}} = \sqrt{\alpha^2-\beta_{\mathfrak{S}}^2} = \alpha$ and $\gamma_{\mathfrak{D}} = \sqrt{\alpha^2-\beta_{\mathfrak{D}}^2} = \sqrt{\alpha^2-1}$, μ can be expressed as a function of the remaining parameters as:

$$\mu = \frac{M}{2}\log\frac{(\rho+1)^2}{2\rho+1} = \log\gamma_{\mathcal{S}}^{2\lambda} - \log\gamma_{\mathfrak{D}}^{2\lambda} . \tag{64}$$

Model (63) corresponds to well-calibrated scores of a PLDA model with isotropic between-over-within class variance ratios ρ_i. In general, the ratios will be different, and the directions with larger ratio will have more impact over the final score. Therefore, rather than fixing λ, we estimate an optimal value from the data. The estimated λ can be interpreted as representing an "effective" number of relevant speaker embedding dimensions. Furthermore, in many practical cases mis-calibration sources can slightly affect the skewness of both distributions. The C-NIG model allows capturing these effects by allowing for non-symmetric target distributions. Since the VΓ skewness can be controlled in a similar way, we relax the model assumptions on $\beta_{\mathcal{D}}$, and we consider the more flexible VΓ model:

$$X|\mathfrak{D} \sim \text{V}\Gamma(\lambda,\alpha,\beta_{\mathfrak{D}},\mu) , \quad X|\mathfrak{S} \sim \text{V}\Gamma(\lambda,\alpha,\beta_{\mathfrak{D}}+1,\mu) . \tag{65}$$

where, in contrast with (63), $\beta_{\mathfrak{D}}$, α and λ are all free parameters of the distribution. In the next section we show that model (65) satisfies the LLR constraint, as long as μ is properly tied to the remaining parameters. Assuming linear calibration as in CMLG and C-NIG, we can write the distribution of observed scores as:

$$S|\mathfrak{S} = \frac{1}{a}(X|\mathfrak{S}) - \frac{b}{a} , \quad S|\mathfrak{D} = \frac{1}{a}(X|\mathfrak{D}) - \frac{b}{a} , \tag{66}$$

and estimate by Maximum Likelihood $\beta_{\mathfrak{D}}$, α and λ, and the calibration parameters a, b. Constrained Variance Gamma (C-VΓ), as C-NIG, is a particular instance of the Constrained GH model. Thus its parameters can be trained as detailed for the GH approach in Sections IV-D and IV-E.

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

C. Generalized Hyperbolic distributions

GH distributions are mean-variance mixtures (57) where the mixing distribution is the 3-parameters Generalized Inverse Gaussian (GIG):

$$f_{\text{GIG}}(x|\lambda,\delta,\gamma) = \left(\frac{\gamma}{\delta}\right)^{\lambda} \frac{x^{\lambda-1}}{2K_\lambda(\delta\gamma)} e^{-\frac{1}{2}\left(\frac{\delta^2}{x}+\gamma^2 x\right)} \ . \quad (67)$$

The resulting GH distribution has 5 parameters, with density:

$$\begin{aligned} f_{\text{GH}}(x|\lambda,\alpha,\beta,\delta,\mu) &= \int \mathcal{N}(x|\mu+\beta v, v) f_{\text{GIG}}(v|\lambda,\delta,\sqrt{\alpha^2-\beta^2}) dv \\ &= \frac{\left(\frac{\gamma}{\delta}\right)^{\lambda} K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)}{\sqrt{2\pi}\alpha^{\lambda-\frac{1}{2}} K_\lambda(\delta\gamma)\left(\sqrt{\delta^2+(x-\mu)^2}\right)^{\frac{1}{2}-\lambda}} e^{\beta(x-\mu)} \ . \end{aligned} \quad (68)$$

where $\alpha > |\beta|, \delta > 0$ and $\gamma = \sqrt{\alpha^2-\beta^2}$. The NIG distribution corresponds to the special case $\lambda = -\frac{1}{2}$:

$$f_{\text{NIG}}(x|\alpha,\beta,\delta\mu) = f_{\text{GH}}\left(x\left|-\frac{1}{2},\alpha,\beta,\delta,\mu\right.\right) \ . \quad (69)$$

The Variance-Gamma distribution corresponds to the boundary case where $\lambda > 0$ and $\delta \to 0$. In this case, the GIG density converges to a Gamma density:

$$f_{GIG}(x|\lambda,\delta,\gamma) \xrightarrow{\delta\to 0} f_\Gamma\left(x\left|\lambda,\frac{\gamma^2}{2}\right.\right) = \frac{\gamma^{2\lambda}}{2^\lambda}\frac{x^{\lambda-1}}{\Gamma(\lambda)} e^{-\frac{1}{2}\gamma^2 x} \ , \quad (70)$$

and the GH density converges to the VΓ density (62)

$$f_{\text{GH}}(x|\lambda,\alpha,\beta,\delta,\mu) \xrightarrow{\delta\to 0} f_{\text{V}\Gamma}(x|\lambda,\alpha,\beta,\mu) \ . \quad (71)$$

In the following we will derive the Constrained GH model using the proper GH density (68). Similar considerations can be readily extended to derive the Constrained Variance Gamma model from the slightly more general GH formulation given in [28], which includes both proper GH (68) and VΓ as special, proper cases. In practice, we implement the VΓ model using proper GH densities with fixed, very small values of δ.

D. Constrained GH model

GH distributions can model well-calibrated scores, provided that the target and non-target distributions are tied. To derive the constraints that well-calibrated GH distributions should satisfy, we assume that non-target scores are generated by a GH density with parameters $(\lambda,\alpha,\beta,\delta,\mu)$:

$$X|\mathfrak{D} \sim \text{GH}(\lambda,\alpha,\beta,\delta,\mu) \ . \quad (72)$$

The MGF of $X|\mathfrak{D}$ is:

$$M_{X|\mathfrak{D}}(t) = e^{\mu t}\frac{\gamma^\lambda}{K_\lambda(\delta\gamma)}\frac{K_\lambda\left(\delta\sqrt{\alpha^2-(\beta+t)^2}\right)}{\left(\sqrt{\alpha^2-(\beta+t)^2}\right)^\lambda} \ , \quad (73)$$

defined for $t \in (-\alpha-\beta, \alpha-\beta)$. The LLR constraint requires $M_{X|\mathfrak{D}}$ to be defined around $t=1$, thus we have to constrain the parameters from $\alpha > |\beta|$ to $\alpha > \max(|\beta|, |\beta+1|)$.The MGF of $X|\mathfrak{S}$ is then given by:

$$M_{X|\mathfrak{S}}(t) = M_{X|\mathfrak{D}}(t+1) = e^{\mu t}\frac{e^\mu\gamma^\lambda}{K_\lambda(\delta\gamma)}\frac{K_\lambda\left(\delta\sqrt{\alpha^2-[(\beta+1)+t]^2}\right)}{\left(\sqrt{\alpha^2-[(\beta+1)+t]^2}\right)^\lambda} \ , \quad (74)$$

with parameters tied to satisfy $M_{X|\mathfrak{S}}(0) = M_{X|\mathfrak{D}}(1) = 1$. Letting $\gamma_{\mathfrak{S}} = \sqrt{\alpha^2-(\beta+1)^2}$:

$$M_{X|\mathfrak{S}}(0) = 1 \iff \frac{e^\mu\gamma^\lambda}{K_\lambda(\delta\gamma)}\frac{K_\lambda(\delta\gamma_{\mathfrak{S}})}{\gamma_{\mathfrak{S}}^\lambda} = 1 \ , \quad (75)$$

and solving for μ:

$$e^\mu = \frac{K_\lambda(\delta\gamma)}{\gamma^\lambda}\frac{\gamma_{\mathfrak{S}}^\lambda}{K_\lambda(\delta\gamma_{\mathfrak{S}})} \ . \quad (76)$$

Letting $\beta_{\mathfrak{S}} = \beta+1$, and replacing (76) in (74) we get the MGF of $X|\mathfrak{S}$:

$$M_{X|\mathfrak{S}}(t) = e^{\mu t}\frac{\gamma_{\mathfrak{S}}^\lambda}{K_\lambda(\delta\gamma_{\mathfrak{S}})}\frac{K_\lambda\left(\delta\sqrt{\alpha^2-(\beta_{\mathfrak{S}}+t)^2}\right)}{\left(\sqrt{\alpha^2-(\beta_{\mathfrak{S}}+t)^2}\right)^\lambda} \ , \quad (77)$$

which is the MGF of a GH distribution with parameters

$$X|\mathfrak{S} \sim GH(\lambda,\alpha,\beta+1,\delta,\mu) \ , \quad (78)$$

Summarizing, the C-GH model assumes that well-calibrated scores are samples of the GH distributions:

$$X|\mathfrak{D} \sim GH(\lambda,\alpha,\beta,\delta,\mu) \ , \quad X|\mathfrak{S} \sim GH(\lambda,\alpha,\beta+1,\delta,\mu) \ , \quad (79)$$

with $\mu = \log\frac{K_\lambda(\delta\gamma)}{\gamma^\lambda}\frac{\gamma_{\mathfrak{S}}^\lambda}{K_\lambda(\delta\gamma_{\mathfrak{S}})}$. Setting $\lambda = -\frac{1}{2}$ gives the C-NIG model of (60), whereas letting $\lambda > 0$ and $\delta \to 0$ leads to the C-VΓ model (65). In this case the parameters constraint becomes $\mu = \log\gamma_{\mathfrak{S}}^{2\lambda} - \log\gamma_{\mathfrak{D}}^{2\lambda}$, which corresponds to (64).

Assuming linear mis-calibration, the observed scores are distributed according to

$$S|\mathfrak{D} = \frac{1}{a}(X|\mathfrak{D}) - \frac{b}{a} \ , \quad S|\mathfrak{S} = \frac{1}{a}(X|\mathfrak{S}) - \frac{b}{a} \ , \quad (80)$$

corresponding to the GH distributions [29]

$$S|\mathfrak{D} \sim GH(\lambda,\overline{\alpha},\overline{\beta}_{\mathfrak{D}},\overline{\delta},\overline{\mu}) \ , \quad S|\mathfrak{S} \sim GH(\lambda,\overline{\alpha},\overline{\beta}_{\mathfrak{S}},\overline{\delta},\overline{\mu}) \quad (81)$$

with

$$\overline{\mu} = \frac{\mu-b}{a}, \ \overline{\alpha} = a\alpha, \ \overline{\beta}_{\mathfrak{D}} = a\beta, \ \overline{\beta}_{\mathfrak{S}} = a\beta+a, \ \overline{\delta} = \frac{\delta}{a} \ . \quad (82)$$

The distributions depend on six free parameters, $\lambda,\overline{\alpha},\overline{\beta_{\mathfrak{D}}},\overline{\beta_{\mathfrak{S}}},\overline{\delta},\overline{\mu}$, which are related to the calibration parameters through (82). The parameters can be estimated by maximizing the weighted log-likelihood

$$\mathscr{L} = \frac{\zeta}{n_{\mathfrak{S}}}\sum_{i\in\mathcal{I}_{\mathfrak{S}}}\log f_{S|\mathfrak{S}}(s_i) + \frac{1-\zeta}{n_{\mathfrak{D}}}\log\sum_{i\in\mathcal{I}_{\mathfrak{D}}} f_{S|\mathfrak{D}}(s_i) \ . \quad (83)$$

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

Let $\overline{\gamma}_{\mathfrak{D}} = \sqrt{\overline{\alpha}^2 - \overline{\beta}_{\mathfrak{D}}^2}$ and $\overline{\gamma}_{\mathfrak{S}} = \sqrt{\overline{\alpha}^2 - \overline{\beta}_{\mathfrak{S}}^2}$. Given λ, $\overline{\alpha}$, $\overline{\beta_{\mathfrak{S}}}$, $\overline{\beta_{\mathfrak{D}}}$, $\overline{\delta}$, $\overline{\mu}$, since $\mu = \log \frac{\gamma_{\mathfrak{S}}^{\lambda} K_{\lambda}(\delta\gamma_{\mathfrak{D}})}{\gamma_{\mathfrak{D}}^{\lambda} K_{\lambda}(\delta\gamma_{\mathfrak{S}})} = \log \frac{\overline{\gamma}_{\mathfrak{S}}^{\lambda} K_{\lambda}(\overline{\delta}\overline{\gamma}_{\mathfrak{D}})}{\overline{\gamma}_{\mathfrak{D}}^{\lambda} K_{\lambda}(\overline{\delta}\overline{\gamma}_{\mathfrak{S}})}$, the calibration parameters can be computed as[5] :

$$a = \overline{\beta}_{\mathfrak{S}} - \overline{\beta}_{\mathfrak{D}} \ , \qquad b = -a\overline{\mu} + \log \frac{\overline{\gamma}_{\mathfrak{S}}^{\lambda} K_{\lambda}\left(\overline{\delta}\overline{\gamma}_{\mathfrak{D}}\right)}{\overline{\gamma}_{\mathfrak{D}}^{\lambda} K_{\lambda}\left(\overline{\delta}\overline{\gamma}_{\mathfrak{S}}\right)} \tag{84}$$

E. Unsupervised C-GH

The unsupervised extension of the C-GH model follows the same strategy of [11] and [10]. Let π denote the target class prior probability. We model observed scores as samples of a random variable S, with density given by:

$$f_S(s) = \pi f_{\mathrm{GH}}(s|\lambda, \overline{\alpha}, \overline{\beta}_{\mathfrak{S}}, \overline{\delta}, \overline{\mu}) + (1-\pi) f_{\mathrm{GH}}(s|\lambda, \overline{\alpha}, \overline{\beta}_{\mathfrak{D}}, \overline{\delta}, \overline{\mu}) \tag{85}$$

i.e. a two-component mixture model whose components are the GH densities of $S|\mathfrak{D}$ and $S|\mathfrak{S}$ in (81). The model parameters can be estimated by maximizing the log-likelihood:

$$\mathscr{L}(\lambda, \overline{\alpha}, \overline{\beta}_{\mathfrak{D}}, \overline{\beta}_{\mathfrak{S}}, \overline{\delta}, \overline{\mu}) = \sum_{i \in \mathcal{I}} \log f_S(s_i) \tag{86}$$

using the EM algorithm, as for CMLG and C-NIG.

V. The EM algorithm for C-GH

Training the C-GH model requires optimizing either (83) or (86) depending on the task. We have observed that general purpose solvers such as L-BFGS [30]–[33], although applicable, often converge to bad local optima of the log-likelihood. We therefore apply an Expectation-Maximization (EM) procedure to estimate the model parameters. For the sake of readability, we drop all the overline symbols, rewriting the model as

$$S|\mathfrak{D} \sim GH(\lambda, \alpha, \beta_{\mathfrak{D}}, \delta, \mu) \ , \quad S|\mathfrak{S} \sim GH(\lambda, \alpha, \beta_{\mathfrak{S}}, \delta, \mu) \tag{87}$$

A. Supervised C-GH

We consider a more general version of the log-likelihood (83), in which each trial score is weighted by factor ζ_i. This will simplify the derivations of the EM steps for the unsupervised model. The objective function is therefore:

$$\mathscr{L} = \sum_{i \in \mathcal{I}_{\mathfrak{S}}} \zeta_i \log f_{S|\mathfrak{S}}(s_i) + \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i \log f_{S|\mathfrak{D}}(s_i) \tag{88}$$

Since the GH distributions are normal mean-variance mixtures, it's natural to consider the mixing R.V. as the hidden variable for the EM algorithm. Let $S = \mu + \beta V + \sqrt{V}Y$, $Y \sim \mathcal{N}(0,1)$, denote a GH-distributed R.V. $S \sim GH(\lambda, \alpha, \beta, \delta, \mu)$. The distribution of S given a value v for V is Gaussian:

$$f_{S|V}(s|v) = \mathcal{N}(s|\mu + \beta v, v) \ . \tag{89}$$

The mixing density is the prior distribution for V:

$$f_V(v) = f_{GIG}\left(v|\lambda, \delta, \sqrt{\alpha^2 - \beta^2}\right) = f_{GIG}\left(v|\lambda, \delta, \gamma\right) \ . \tag{90}$$

[5]Alternatively, the calibration transformation can be computed directly from the GH densities $x = f_{cal}(s) = as + b = \log \frac{f_{GH}(s|\lambda, \overline{\alpha}, \overline{\beta}_{\mathfrak{S}}, \overline{\delta}, \overline{\mu})}{f_{GH}(s|\lambda, \overline{\alpha}, \overline{\beta}_{\mathfrak{D}}, \overline{\delta}, \overline{\mu})}$.

The joint distribution of S and V is therefore given by:

$$\begin{aligned} \log f_{S,V}(s,v) = c - \frac{1}{2}\log v - \frac{1}{2}\frac{(s-\mu-\beta v)^2}{v} + \lambda \log \gamma \\ - \lambda \log \delta + (\lambda - 1)\log v - \log K_{\lambda}(\delta\gamma) - \frac{1}{2}\frac{\delta^2}{v} - \frac{1}{2}\gamma^2 v \ , \end{aligned} \tag{91}$$

where c collects all constant terms. The posterior distribution for V given S is also a GIG [27]:

$$f_{V|S}(v|s) = f_{\mathrm{GIG}}\left(v \,\middle|\, \lambda - \frac{1}{2}, \sqrt{\delta^2 + (s-\mu)^2}, \alpha\right) \ . \tag{92}$$

It is worth noting that the posterior distribution does not depend on parameter β, thus it has the same shape both if s_i is a target or it is a non-target score. Therefore, we will denote the posterior distributions for both target and non-target scores as $f_{V|S}$ rather than $f_{V|S,\mathfrak{S}}$ and $f_{V|S,\mathfrak{D}}$.

The EM auxiliary function is given by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i \mathbb{E}_{f_{V|S}(v_i|s_i, \boldsymbol{\theta}^*)}\left[\log f_{S,V|\mathfrak{D}}(s_i, v_i|\boldsymbol{\theta})\right] \\ + \sum_{i \in \mathcal{I}_{\mathfrak{S}}} \zeta_i \mathbb{E}_{f_{V|S}(s_i|v_i, \boldsymbol{\theta}^*)}\left[\log f_{S,V|\mathfrak{S}}(s_i, v_i|\boldsymbol{\theta})\right] \end{aligned} \tag{93}$$

where $\boldsymbol{\theta} = (\lambda, \alpha, \beta_{\mathfrak{D}}, \beta_{\mathfrak{S}}, \delta, \mu)$ is the set of parameters to estimate, $\boldsymbol{\theta}^* = (\lambda^*, \alpha^*, \beta_{\mathfrak{D}}^*, \beta_{\mathfrak{S}}^*, \delta^*, \mu^*)$ is the current set of parameters, and

$$\begin{aligned} f_{S,V|\mathfrak{D}}(s,v) &= \mathcal{N}(s|\mu + \beta_{\mathfrak{D}} v, v) \cdot f_{\mathrm{GIG}}(v|\lambda, \delta, \gamma_{\mathfrak{D}}) \\ f_{S,V|\mathfrak{S}}(s,v) &= \mathcal{N}(s|\mu + \beta_{\mathfrak{S}} v, v) \cdot f_{\mathrm{GIG}}(v|\lambda, \delta, \gamma_{\mathfrak{S}}) \end{aligned} \tag{94}$$

The expectations in (93) can be computed as:

$$\begin{aligned} \mathbb{E}_{f_{V|S}(s|v,\boldsymbol{\theta}^*)}\left[\log f_{S|V,\mathfrak{D}}(s|v,\boldsymbol{\theta}) f_{V|\mathfrak{D}}(v|\boldsymbol{\theta})\right] = c_1 + \mu s \mathbb{E}\left[\frac{1}{v}\right] \\ - \frac{1}{2}\mu^2 \mathbb{E}\left[\frac{1}{v}\right] + \beta_{\mathfrak{D}} s - \beta_{\mathfrak{D}}\mu + \lambda \log \gamma_{\mathfrak{D}} - \lambda \log \delta \\ + \lambda \mathbb{E}\left[\log v\right] - \log K_{\lambda}(\delta\gamma_{\mathfrak{D}}) - \frac{1}{2}\delta^2 \mathbb{E}\left[\frac{1}{v}\right] - \frac{1}{2}\alpha^2 \mathbb{E}\left[v\right] \ , \end{aligned} \tag{95}$$

where c_1 collects all terms that do not depend on the model parameters $\boldsymbol{\theta}$. The expressions for the target class $\mathbb{E}_{f_{V|S}(s|v,\boldsymbol{\theta}^*)}\left[\log f_{S|V,\mathfrak{S}}(s|v,\boldsymbol{\theta}) f_{V|\mathfrak{S}}(v|\boldsymbol{\theta})\right]$ are obtained simply replacing $\gamma_{\mathfrak{D}}$ by $\gamma_{\mathfrak{S}}$, and $\beta_{\mathfrak{D}}$ by $\beta_{\mathfrak{S}}$ in (95). We make use of well known results for GIG distributions [34] to compute the expectations. Let

$$\phi^*(s) = \sqrt{\delta^{*2} + (s - \mu^*)^2} \ . \tag{96}$$

Recalling the posterior distribution for $V|S$ in (92), the expectations are given by:

$$\begin{aligned} \mathbb{E}_{f_{V|S}(v|s,\boldsymbol{\theta}^*)}\left[v\right] &= \frac{\phi^*(s) K_{\lambda^*+\frac{1}{2}}\left(\alpha^*\phi^*(s)\right)}{\alpha K_{\lambda^*-\frac{1}{2}}\left(\alpha^*\phi^*(s)\right)} \\ \mathbb{E}_{f_{V|S}(v|s,\boldsymbol{\theta}^*)}\left[\frac{1}{v}\right] &= \frac{\alpha K_{\lambda^*+\frac{1}{2}}\left(\alpha^*\phi^*(s)\right)}{\phi^*(s) K_{\lambda^*-\frac{1}{2}}\left(\alpha^*\phi^*(s)\right)} - \frac{2\lambda^* - 1}{\left[\phi^*(s)\right]^2} \\ \mathbb{E}_{f_{V|S}(v|s,\boldsymbol{\theta}^*)}\left[\log v\right] &= \log \frac{\phi^*(s)}{\alpha} + \left(\partial_{\nu}\left[\log K_{\nu}\right]\right)_{\lambda^*-\frac{1}{2}}\left(\alpha\phi^*(s)\right) \end{aligned} \tag{97}$$

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

where $(\partial_\nu [\log K_\nu])_{\lambda^* - \frac{1}{2}}(x)$ denotes the derivative of the logarithm of the Bessel function $[\log K_\nu(x)]$ with respect to the order ν, evaluated at $\nu = \lambda^* - \frac{1}{2}$. Since we are not aware of simple expressions for $\partial_\nu [\log K_\nu]$, in this work we approximate this derivative through finite differences. Replacing the expectations in (93), and applying some algebraic manipulations, we can express the auxiliary function in terms of class-dependent statistics:

$$\begin{aligned}
\mathcal{Z}_{\mathfrak{G}} &= \sum_{i \in \mathcal{I}_{\mathfrak{G}}} \zeta_i \ , & \mathcal{Z}_{\mathfrak{D}} &= \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i \ , \\
\mathcal{F}_{\mathfrak{G}} &= \sum_{i \in \mathcal{I}_{\mathfrak{G}}} \zeta_i s_i \ , & \mathcal{F}_{\mathfrak{D}} &= \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i s_i \ , \\
\mathcal{V}_{\mathfrak{G}} &= \sum_{i \in \mathcal{I}_{\mathfrak{G}}} \zeta_i \mathbb{E}\left[v_i\right] \ , & \mathcal{V}_{\mathfrak{D}} &= \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i \mathbb{E}\left[v_i\right] \ , \\
\mathcal{W}_{\mathfrak{G}} &= \sum_{i \in \mathcal{I}_{\mathfrak{G}}} \zeta_i \mathbb{E}\left[\frac{1}{v_i}\right] \ , & \mathcal{W}_{\mathfrak{D}} &= \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i \mathbb{E}\left[\frac{1}{v_i}\right] \ , \\
\mathcal{G}_{\mathfrak{G}} &= \sum_{i \in \mathcal{I}_{\mathfrak{G}}} \zeta_i s_i \mathbb{E}\left[\frac{1}{v_i}\right] \ , & \mathcal{G}_{\mathfrak{D}} &= \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i s_i \mathbb{E}\left[\frac{1}{v_i}\right] \ , \\
\mathcal{L}_{\mathfrak{G}} &= \sum_{i \in \mathcal{I}_{\mathfrak{G}}} \zeta_i \mathbb{E}\left[\log v_i\right] \ , & \mathcal{L}_{\mathfrak{D}} &= \sum_{i \in \mathcal{I}_{\mathfrak{D}}} \zeta_i \mathbb{E}\left[\log v_i\right] \ ,
\end{aligned} \tag{98}$$

and global statistics:

$$\begin{aligned}
\mathcal{Z} &= \mathcal{Z}_{\mathfrak{G}} + \mathcal{Z}_{\mathfrak{D}} \ , & \mathcal{W} &= \mathcal{W}_{\mathfrak{G}} + \mathcal{W}_{\mathfrak{D}} \ , \\
\mathcal{G} &= \mathcal{G}_{\mathfrak{G}} + \mathcal{G}_{\mathfrak{D}} \ , & \mathcal{L} &= \mathcal{L}_{\mathfrak{G}} + \mathcal{L}_{\mathfrak{D}} \ .
\end{aligned} \tag{99}$$

The auxiliary function is given by:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = c_2 + \mathcal{G}\mu - \frac{1}{2}\mathcal{W}\mu^2 + \mathcal{F}_{\mathfrak{D}}\beta_{\mathfrak{D}} + \mathcal{F}_{\mathfrak{G}}\beta_{\mathfrak{G}} - \mathcal{Z}_{\mathfrak{D}}\mu\beta_{\mathfrak{D}} \\
- \mathcal{Z}_{\mathfrak{G}}\mu\beta_{\mathfrak{G}} + \mathcal{Z}_{\mathfrak{D}}\lambda\log\gamma_{\mathfrak{D}} + \mathcal{Z}_{\mathfrak{G}}\lambda\log\gamma_{\mathfrak{G}} - \mathcal{Z}\lambda\log\delta + \mathcal{L}\lambda \\
- \mathcal{Z}_{\mathfrak{D}}\log K_\lambda(\delta\gamma_{\mathfrak{D}}) - \mathcal{Z}_{\mathfrak{G}}\log K_\lambda(\delta\gamma_{\mathfrak{G}}) - \frac{1}{2}\mathcal{W}\delta^2 - \frac{1}{2}\mathcal{V}\alpha^2 \ ,
\end{aligned} \tag{100}$$

where c_2 collects all terms that do not depend on the parameters $\boldsymbol{\theta}$, and $\gamma_{\mathfrak{D}}$ and $\gamma_{\mathfrak{G}}$ are functions of α, $\beta_{\mathfrak{D}}$ and $\beta_{\mathfrak{G}}$.

Maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ with respect to $\boldsymbol{\theta}$ cannot be performed in closed form. We thus rely on numerical optimization to estimate the updated parameters in the M-step. However, for some parameters the domain is not $\mathbb{R}$. The constraints that should be enforced are $\alpha > \sqrt{\max(\beta_{\mathfrak{D}}^2, \beta_{\mathfrak{G}}^2)}$, $\beta_{\mathfrak{G}} > \beta_{\mathfrak{D}}$ and $\delta > 0$. Furthermore, the complexity of the computation of Bessel functions is proportional to the order λ of the function itself. We have shown in section IV that the effective NIG approximation corresponds to $\lambda = -\frac{1}{2}$. We have also shown that, for the C-VG model, we can interpret λ as the number of "effective" input dimensions. Therefore, we add an additional (not required by the model) constraint to bound the value of λ to $\lambda \in (-\frac{M}{2}, \frac{M}{2})$[6], where M is the input dimensionality. To avoid using numerical optimizers with explicit constraints, we re-parametrize the objective function expressing the distribution parameters $(\lambda, \alpha, \beta_{\mathfrak{D}}, \beta_{\mathfrak{G}}, \delta)$ as an invertible function of the optimization parameters $(\hat{\lambda}, \Delta_\alpha, \beta_0, \Delta_\beta^{\log}, \delta^{\log})$:

$$\begin{aligned}
&\beta_{\mathfrak{D}} = \beta_0 - \frac{1}{2}e^{\Delta_\beta^{\log}} \ , \quad \beta_{\mathfrak{G}} = \beta_0 + \frac{1}{2}e^{\Delta_\beta^{\log}} \ , \quad \delta = e^{\delta^{\log}} \\
&\alpha = \sqrt{\sigma_k(\beta_{\mathfrak{G}}^2, \beta_{\mathfrak{D}}^2) + \Delta_\alpha^2} \ , \quad \lambda = \frac{M}{2}\tanh(\hat{\lambda}) \ ,
\end{aligned} \tag{101}$$

where $\sigma_k(a, b)$ is the log-sum-exp or softmax[7] function[8]

$$\sigma_k(a, b) = \frac{1}{k}\log\left(e^{ka} + e^{kb}\right) \tag{102}$$

where k controls the smoothness of the softmax function (in our experiments we found $k = 1$ to be effective). Maximization of the auxiliary function is based on the L-BFGS [30]–[33] algorithm. The algorithm requires computing the gradients with respect to the optimization parameters $(\hat{\lambda}, \Delta_\alpha, \beta_0, \Delta_\beta^{\log}, \mu, \delta^{\log})$. These gradients can be computed from the derivatives of Q w.r.t. the original parameters. Recalling the derivative of the log-Bessel function:

$$D_\lambda(x) = \frac{\partial \log K_\lambda(x)}{\partial x} = \frac{\lambda}{x} - \frac{K_{\lambda+1}(x)}{K_\lambda(x)} = -\frac{K_{\lambda-1}(x)}{K_\lambda(x)} - \frac{\lambda}{x} \ , \tag{103}$$

the partial derivatives of Q are:

$$\begin{aligned}
&\frac{\partial Q}{\partial \gamma_{\mathfrak{D}}} = \mathcal{Z}_{\mathfrak{D}}\delta\frac{K_{\lambda+1}(\delta\gamma_{\mathfrak{D}})}{K_\lambda(\delta\gamma_{\mathfrak{D}})} \ , \quad \frac{\partial Q}{\partial \gamma_{\mathfrak{G}}} = \mathcal{Z}_{\mathfrak{G}}\delta\frac{K_{\lambda+1}(\delta\gamma_{\mathfrak{G}})}{K_\lambda(\delta\gamma_{\mathfrak{G}})} \ , \\
&\frac{\partial Q}{\partial \beta_{\mathfrak{D}}} = \mathcal{F}_{\mathfrak{D}} - \mathcal{Z}_{\mathfrak{D}}\mu - \frac{\beta_{\mathfrak{D}}}{\gamma_{\mathfrak{D}}}\frac{\partial Q}{\partial \gamma_{\mathfrak{D}}} \ , \quad \frac{\partial Q}{\partial \beta_{\mathfrak{G}}} = \mathcal{F}_{\mathfrak{G}} - \mathcal{Z}_{\mathfrak{G}}\mu - \frac{\beta_{\mathfrak{G}}}{\gamma_{\mathfrak{G}}}\frac{\partial Q}{\partial \gamma_{\mathfrak{G}}} \ , \\
&\frac{\partial Q}{\partial \alpha} = \frac{\alpha}{\gamma_{\mathfrak{D}}}\frac{\partial Q}{\partial \gamma_{\mathfrak{D}}} + \frac{\alpha}{\gamma_{\mathfrak{G}}}\frac{\partial Q}{\partial \gamma_{\mathfrak{G}}} - \alpha\mathcal{V} \ , \\
&\frac{\partial Q}{\partial \delta} = -\mathcal{Z}_{\mathfrak{D}}\gamma_{\mathfrak{D}}\frac{K_{\lambda-1}(\delta\gamma_{\mathfrak{D}})}{K_\lambda(\delta\gamma_{\mathfrak{D}})} - \mathcal{Z}_{\mathfrak{G}}\gamma_{\mathfrak{G}}\frac{K_{\lambda-1}(\delta\gamma_{\mathfrak{G}})}{K_\lambda(\delta\gamma_{\mathfrak{G}})} - \delta\mathcal{W} \ , \\
&\frac{\partial Q}{\partial \mu} = \mathcal{G} - \mathcal{W}\mu - \mathcal{Z}_{\mathfrak{D}}\beta_{\mathfrak{D}} - \mathcal{Z}_{\mathfrak{G}}\beta_{\mathfrak{G}} \ ,
\end{aligned} \tag{104}$$

The derivative $\frac{\partial Q}{\partial \lambda}$ requires differentiating the Bessel function w.r.t. its order, and was therefore approximated by finite differences. The derivatives with respect to the optimization parameters $(\hat{\lambda}, \Delta_\alpha, \beta_0, \Delta_\beta^{\log}, \delta^{\log})$ can be obtained by applying the chain rule, and are given by:

$$\begin{aligned}
&\frac{\partial Q}{\partial \Delta_\alpha} = \frac{\partial Q}{\partial \alpha}\frac{\Delta_\alpha}{\alpha} \ , \ \frac{\partial Q}{\partial \delta^{\log}} = \frac{\partial Q}{\partial \delta}e^{\delta} \ , \ \frac{\partial Q}{\partial \hat{\lambda}} = \frac{\partial Q}{\partial \lambda}\frac{M}{2}\left(1 - \tanh^2(\lambda)\right) \ , \\
&\frac{\partial Q}{\partial \Delta_\beta^{\log}} = \sum_{\mathfrak{h} \in \{\mathfrak{G}, \mathfrak{D}\}} \left(\frac{\partial Q}{\partial \alpha}\frac{\beta_{\mathfrak{h}}}{\alpha}\frac{e^{k\beta_{\mathfrak{h}}}}{e^{k\beta_{\mathfrak{D}}} + e^{k\beta_{\mathfrak{G}}}} + \frac{\partial Q}{\partial \beta_{\mathfrak{h}}}\right)\left(\beta_{\mathfrak{h}} - \beta_0\right) \ , \\
&\frac{\partial Q}{\partial \beta_0} = \sum_{\mathfrak{h} \in \{\mathfrak{G}, \mathfrak{D}\}} \left(\frac{\partial Q}{\partial \alpha}\frac{\beta_{\mathfrak{h}}}{\alpha}\frac{e^{k\beta_{\mathfrak{h}}}}{e^{k\beta_{\mathfrak{D}}} + e^{k\beta_{\mathfrak{G}}}} + \frac{\partial Q}{\partial \beta_{\mathfrak{h}}}\right) \ ,
\end{aligned} \tag{105}$$

Although more effective than numerical optimization, the EM algorithm exhibited, in our experiments, slow convergence. We therefore applied a Quasi-Newton (QN) acceleration scheme [35]. The method combines the log-likelihood gradient with the ascent direction given by the M-step solution. The log-likelihood gradient can be computed from the auxiliary

[6]The densities of the C-VΓ model would require $\lambda > 0$. Since we implement C-VΓ as a C-GH model with small δ, we can relax the constraint $\lambda > 0$. In practice, we did not obtain estimated values $\lambda \leq 0$ in our tests.

[7]The name softmax is commonly used to refer to the multivariate function $\frac{e^{x_i}}{\sum_i e^{x_i}}$, which, however, does not correspond to a soft maximum, but rather to a soft $\arg\max$. The soft $\arg\max$ is the gradient of the softmax function in (102)

[8]We replace the maximum with a smooth maximum function to keep the re-parametrization differentiable.

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

function derivatives or, alternatively, from the derivatives of the logarithm of a GH density $\log f_{\text{GH}}(x|\mu,\alpha,\beta,\delta,\lambda)$:

$$\begin{aligned}
\frac{\partial \log f_{\text{GH}}}{\partial \alpha} &= \frac{\alpha\lambda}{\gamma^2} - \frac{\alpha\delta}{\gamma} D_{\lambda}(\delta\gamma) + \phi(x) D_{\lambda-\frac{1}{2}}\left(\alpha\phi(x)\right) - \frac{\lambda-\frac{1}{2}}{\alpha} \ , \\
\frac{\partial \log f_{\text{GH}}}{\partial \beta} &= \frac{\delta\beta}{\gamma} D_{\lambda}(\delta\gamma) - \frac{\lambda\beta}{\gamma^2} + x - \mu \ , \\
\frac{\partial \log f_{\text{GH}}}{\partial \delta} &= \frac{\alpha\delta}{\phi(x)} D_{\lambda-\frac{1}{2}}\left(\alpha\phi(x)\right) - \gamma D_{\lambda}\left(\delta\gamma\right) - \frac{\lambda}{\delta} + \left(\lambda-\frac{1}{2}\right)\frac{\delta}{\phi(x)^2} \ , \\
\frac{\partial \log f_{\text{GH}}}{\partial \mu} &= \left(\frac{\alpha}{\phi(x)} D_{\lambda}\left(\alpha\phi\left(x\right)\right) + \frac{\lambda-\frac{1}{2}}{\phi(x)^2}\right)(\mu - x) - \beta \ ,
\end{aligned} \tag{106}$$

where $\phi(x) = \sqrt{\delta^2 + (x-\mu)^2}$. As for the EM auxiliary function, the derivative w.r.t. λ requires differentiating the Bessel function with respect to its order, which we approximate via finite differences.

B. Unsupervised C-GH

We introduce a latent variable $\mathcal{H}$ for class labels, with Bernoulli prior distribution

$$P_{\mathcal{H}}(\mathfrak{S}) = \pi \ , \quad P_{\mathcal{H}}(\mathfrak{D}) = 1-\pi \ . \tag{107}$$

Given $\mathcal{H}$, the conditional distributions for target and non-target scores are the same as in the supervised model.

The complete data log-likelihood can be expressed as[9]

$$f_{S,V,\mathcal{H}}(s,v,\mathfrak{h}) = f_{S|V,\mathfrak{h}}(s|v) f_{V|\mathfrak{h}}(v) P_{\mathcal{H}}(\mathfrak{h}) \tag{108}$$

where $\mathfrak{h} \in \{\mathfrak{D},\mathfrak{S}\}$. The marginal distribution for the score is:

$$\begin{aligned}
f_S(s) &= \sum_{\mathfrak{h}\in\{\mathfrak{D},\mathfrak{S}\}} P_{\mathcal{H}}(\mathfrak{h}) \int_v f_{S|V,\mathfrak{h}}(s|v) f_{V|\mathfrak{h}}(v) dv \\
&= \pi f_{S|\mathfrak{S}}(s) + (1-\pi) f_{S|\mathfrak{D}}(s)
\end{aligned} \tag{109}$$

where

$$f_{S|\mathfrak{h}}(s) = f_{\text{GH}}(\lambda,\alpha,\beta_{\mathfrak{h}},\delta,\mu) \ , \tag{110}$$

and the log-likelihood for the observed scores is:

$$\mathscr{L} = \sum_{i\in\mathcal{I}} \log\left[\pi f_{S|\mathfrak{S}}(s_i) + (1-\pi) f_{S|\mathfrak{D}}(s_i)\right] \tag{111}$$

The EM algorithm allows computing ML estimates of the distribution parameters. In contrast with the supervised case, the latent variables include both V and $\mathcal{H}$. The prior distribution for V and $\mathcal{H}$ can be factorized as

$$f_{V,\mathcal{H}}(v,\mathfrak{h}) = f_{V|\mathfrak{h}}(v) P_{\mathcal{H}}(\mathfrak{h}) \ . \tag{112}$$

The EM algorithm requires maximizing the auxiliary function

$$Q(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}}^*) = \sum_{i\in\mathcal{I}} \mathbb{E}_{f_{V,\mathcal{H}|S}(v_i,\mathfrak{h}_i|s_i,\overline{\boldsymbol{\theta}}^*)}\left[\log f_{S,V,\mathcal{H}}(s_i,v_i,\mathfrak{h}_i|\overline{\boldsymbol{\theta}})\right] \ , \tag{113}$$

where we made explicit the optimization parameters $\overline{\boldsymbol{\theta}} = (\lambda,\alpha,\beta_{\mathfrak{D}},\beta_{\mathfrak{S}},\delta,\mu,\pi)$ and the current parameter estimate $\overline{\boldsymbol{\theta}}^* = (\lambda^*,\alpha^*,\beta^*_{\mathfrak{D}},\beta^*_{\mathfrak{S}},\delta^*,\mu^*,\pi^*)$. The posterior distribution required for the E-step can be expressed as

$$f_{V,\mathcal{H}|S}(v,\mathfrak{h}|s,\overline{\boldsymbol{\theta}}^*) = f_{V|S}(v|s,\overline{\boldsymbol{\theta}}^*) P_{\mathcal{H}|S}(\mathfrak{h}|s,\overline{\boldsymbol{\theta}}^*) \tag{114}$$

[9]We denote densities conditioned on values of $\mathcal{H}=\mathfrak{h}$ as $f_{\cdot|\mathfrak{h}}(\cdot)$ rather than $f_{\cdot|\mathcal{H}}(\cdot|\mathfrak{h})$. The choice allows keeping the notation of previous sections for target and non-target densities (e.g. $f_{\cdot|\mathfrak{S}}(\cdot)$ rather than $f_{\cdot|\mathcal{H}}(\cdot|\mathfrak{S})$)

where $f_{V|S}$ does not depend on $\mathfrak{h}$ (Section V-A) and $P_{\mathcal{H}|S}(\mathfrak{h}|s,\overline{\boldsymbol{\theta}}^*)$ can be computed as

$$P_{\mathcal{H}|S}(\mathfrak{h}|s,\overline{\boldsymbol{\theta}}^*) = \frac{f_{S|\mathfrak{h}}(s|\overline{\boldsymbol{\theta}}^*)}{\pi^* f_{S|\mathfrak{S}}(s|\overline{\boldsymbol{\theta}}^*) + (1-\pi^*) f_{S|\mathfrak{D}}(s|\overline{\boldsymbol{\theta}}^*)} \ . \tag{115}$$

Let $\boldsymbol{\theta}^* = (\lambda^*,\alpha^*,\beta^*_{\mathfrak{D}},\beta^*_{\mathfrak{S}},\delta^*,\mu^*)$ be the current GH parameters estimate. The auxiliary function can be rewritten as

$$\begin{aligned}
Q(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}}^*) = \sum_{i\in\mathcal{I}} \mathbb{E}_{P_{\mathcal{H}|S}(\mathfrak{h}_i|s_i,\overline{\boldsymbol{\theta}}^*)} \Bigg[& \mathbb{E}_{f_{V|S}(v_i|s_i,\boldsymbol{\theta}^*)} \\
& \left[\log f_{S,V|\mathfrak{h}_i}(s_i,v_i|\overline{\boldsymbol{\theta}}) + \log P_{\mathcal{H}}(\mathfrak{h}_i|\overline{\boldsymbol{\theta}})\right] \Bigg] \ .
\end{aligned} \tag{116}$$

Let $\hat{\zeta}_i = P_{\mathcal{H}|S}(\mathfrak{S}|s_i,\overline{\boldsymbol{\theta}}^*)$, so that $P_{\mathcal{H}|S}(\mathfrak{D}|s_i,\overline{\boldsymbol{\theta}}^*) = 1-\hat{\zeta}_i$. The auxiliary function can be written as the sum of two terms, the first depending only on the GH parameters $\boldsymbol{\theta} = (\lambda,\alpha,\beta_{\mathfrak{D}},\beta_{\mathfrak{S}},\delta,\mu)$, and the second only on the weight π:

$$Q(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}}^*) = Q_1(\boldsymbol{\theta},\overline{\boldsymbol{\theta}}^*) + Q_2(\pi,\overline{\boldsymbol{\theta}}^*) \tag{117}$$

with

$$\begin{aligned}
Q_1(\boldsymbol{\theta},\overline{\boldsymbol{\theta}}^*) = & \sum_{i\in\mathcal{I}} \hat{\zeta}_i \mathbb{E}_{f_{V|S}(v_i|s_i,\boldsymbol{\theta}^*)}\left[\log f_{S,V|\mathfrak{S}}(s_i,v_i|\boldsymbol{\theta})\right] \\
& + \sum_{i\in\mathcal{I}} \left(1-\hat{\zeta}_i\right) \mathbb{E}_{f_{V|S}(v_i|s_i,\boldsymbol{\theta}^*)}\left[\log f_{S,V|\mathfrak{D}}(s_i,v_i|\boldsymbol{\theta})\right] \ , \\
Q_2(\pi,\overline{\boldsymbol{\theta}}^*) = & \sum_{i\in\mathcal{I}} \hat{\zeta}_i \log\pi + \sum_{i\in\mathcal{I}} \left(1-\hat{\zeta}_i\right)\log(1-\pi) \ .
\end{aligned} \tag{118}$$

We can thus independently optimize Q_1 and Q_2. We observe that $Q_1(\boldsymbol{\theta},\overline{\boldsymbol{\theta}}^*)$ has a very similar expression as the auxiliary function for the supervised task (93). In particular, we can interpret Q_1 as the auxiliary function of a supervised task, where the samples are considered as belonging to both the target and non-target class with weights $\hat{\zeta}_i$ and $1-\hat{\zeta}_i$, respectively. $Q_1(\boldsymbol{\theta},\overline{\boldsymbol{\theta}}^*)$ can be expressed as in (100), provided that the statistics (98) are computed taking the summations over all samples, and using weights $\hat{\zeta}_i$ for the summations of statistics with subscript $\mathfrak{S}$, and weight $1-\hat{\zeta}_i$ for the statistics with subscript $\mathfrak{D}$. For example, the zero and first order statistics are obtained as:

$$\begin{aligned}
&\mathcal{Z}_{\mathfrak{S}} = \sum_{i\in\mathcal{I}} \hat{\zeta}_i \ , \quad \mathcal{Z}_{\mathfrak{S}} = \sum_{i\in\mathcal{I}} \left(1-\hat{\zeta}_i\right) \ , \quad \mathcal{Z} = \mathcal{Z}_{\mathfrak{S}} + \mathcal{Z}_{\mathfrak{D}} \\
&\mathcal{F}_{\mathfrak{S}} = \sum_{i\in\mathcal{I}} \hat{\zeta}_i s_i \ , \quad \mathcal{F}_{\mathfrak{D}} = \sum_{i\in\mathcal{I}} \left(1-\hat{\zeta}_i\right) s_i \ ,
\end{aligned} \tag{119}$$

and the same applies to the other terms. The M-step for Q_1 can be implemented as the M-step for the supervised auxiliary function. The second term $Q_2(\pi,\overline{\boldsymbol{\theta}}^*)$ can be optimized in closed form, with the optimum given by:

$$\pi = \frac{\mathcal{Z}_{\mathfrak{S}}}{\mathcal{Z}} \ . \tag{120}$$

Since the QN approach cannot enforce the constraint $\pi \in (0,1)$, we re-parametrize π as $\pi = \frac{1}{1+e^{-\omega}}$, and optimize the log-likelihood w.r.t. the parameter ω.

VI. Experimental results

In this section we analyze the performance of the proposed models for supervised and unsupervised tasks with different speaker vector front-ends and classification back-ends.

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

A. Classifiers

The considered back-ends are PLDA and two PLDA-derived classifiers: Non-Linear PLDA (NL-PLDA) [21], [22] and Pairwise Support Vector Machine (PSVM) [23], [24].

1) NL-PLDA: NL-PLDA extends PLDA introducing a non-linear, invertible function aimed at transforming the speaker vectors so that their distribution better matches the PLDA assumptions, improving accuracy and reducing mis-calibration effects due to inaccurate modeling assumptions.

2) Pairwise SVM: 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. The resulting model outperforms PLDA in several benchmarks. Since PSVM and PLDA scoring functions have the same formal expression, we expect our approach to be effective also for the PSVM back-end.

B. Benchmarks

We consider three different datasets: SRE 2019 [15], SRE 2012 [16] and SRE 2010 [17], each based on a different front-end. For the unsupervised scenario, we consider different percentages of target and non-target priors. However, rather than sampling different amounts of non-target trials, we keep the same training set, but we differently re-weight target and non-target samples. This allows comparing models trained using different proportions of trials.

1) SRE 2019: The SRE 2019 front-end consists of a Deep Neural Network (DNN) with the same topology as in [36]. The DNN input consists of 24-dimensional Perceptual Linear Predictors (PLP) features, and the speaker embeddings are 512-dimensional. The speaker embeddings have been processed by means of LDA, which reduces the dimensionality to 400 for PSVM and to 150 for PLDA and NL-PLDA, followed by whitening. For PLDA, we consider both length-normalized and raw embeddings. The NL-PLDA backend can incorporate an a utterance-dependent scaling factor similar to length normalization as detailed in [21], [22]. We test the model both with and without this scaling factor. Within-Class Covariance (WCCN) normalization was applied to length-normalized embeddings for PSVM training. Since typical applications of unsupervised calibration involve development data that closely mimics the test population, 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 the hybrid GMM/DNN framework [37]–[39]. 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 [40]. Overall, the UBM has 2048 components. The speaker vectors are obtained from a 400-dimensional e-vector extractor [19]. The NL-PLDA model has been trained on whitened e-vectors, and includes an utterance-dependent scaling factor [21], [22], whereas length normalization followed by WCCN was applied for the PSVM back-end. Tests were performed on the extended tel-tel core condition (condition 5). The test sets were divided in two, non overlapping, parts. The first part, that 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 backend is a PLDA classifier. I-vectors pre-processing consists in whitening and length normalization. The tests were performed on the female extended tel-tel condition (condition 5). The test set was divided in two, non overlapping, parts. The first part, that comprises 25% of the enrollment segments, was used to estimate the calibration parameters. The remaining part was used as evaluation.

C. Evaluation Metrics

Results are reported in terms of Cost of Log-Likelihood Ratio C_{llr} [4], [41], [42] and of the primary metric C_{prim} defined for each task. In several real applications low False Acceptance (FA) operating points are far more interesting than low False Rejection (FR) operating points. Since C_{llr} does not differentiate from bad calibration in low FA and low FR regions, we also propose a metric that computes the re-normalized contribution to C_{llr} due to applications where the FA cost c_{fa} is larger than the FR cost c_{fr}. Rather than assuming a uniform prior over $[0, 1]$ for the parameter that represents the range of possible applications $t = \frac{c_{fa}}{c_{fa}+c_{fr}}$ as for C_{llr} [4], [41], [42], we assume a uniform prior over $[0.5, 1]$. We denote this measure as C^{ℓ}_{llr} [10] We also provide normalized Bayes error rate [41] plots for some of the systems. These graphs show the normalized Detection Cost Function (DCF) corresponding to different target prior log-odds $x = \log \frac{p}{1-p}$, where p is a synthetic prior.

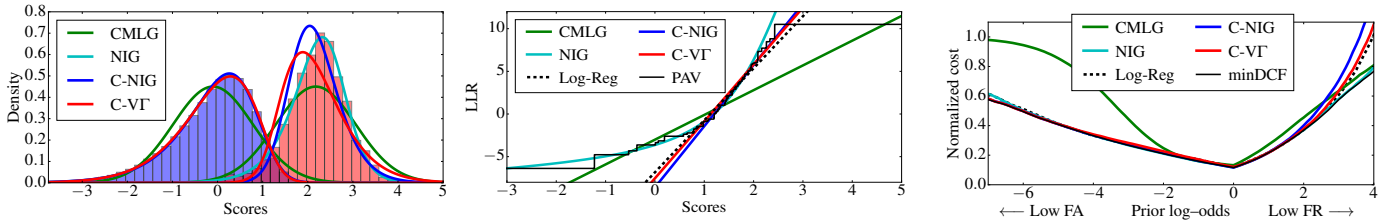
D. Baseline systems and model initialization

For supervised scenarios the baseline systems are Log-Reg, CMLG, and the unconstrained NIG model of [8]. For the unsupervised scenario, the baselines are an unsupervised CMLG [11] model, and the unsupervised extension of the unconstrained NIG model [10]. To avoid possible bias due to poor initialization, the unsupervised CMLG and NIG models have been initialized using oracle values obtained from the corresponding supervised models. For C-NIG and C-VΓ we investigate both the oracle and a completely unsupervised initialization, consisting in the following steps: (i) scores are centered and whitened to avoid possible numerical issues due to score dynamics; (ii) a single NIG or VΓ distribution is fitted over all scores, and the resulting parameters are used as initial estimates for the tied parameters and for $\beta_{\mathfrak{D}}$; (iii) calibration parameters are set to $a = 1$ and $b = 0$, so that initial value for the target distribution skewness becomes $\beta_{\mathfrak{S}} = \beta_{\mathfrak{D}} + 1$. Oracle-initialized models are denoted by a symbol * in the following tables and plots.

[10]The corresponding measure for low FR regions, C^{r}_{llr}, can be obtained assuming uniform prior for t over $[0, 0.5]$. C_{llr} is then given by the average of C^{l}_{llr} and the C^{r}_{llr}.

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

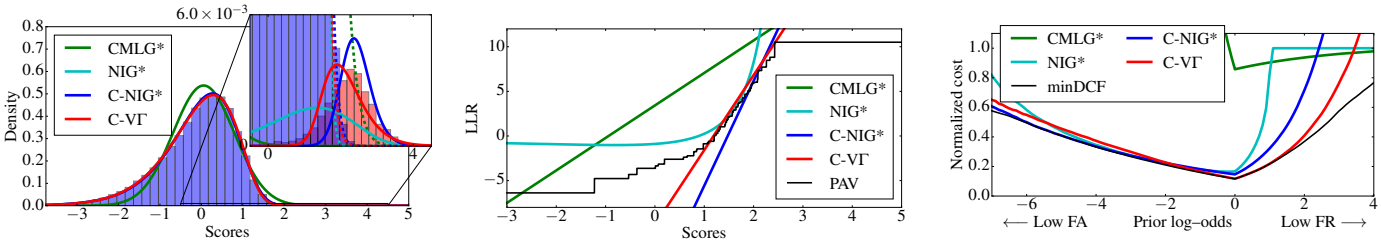


(a) Normalized target and non-target score histograms, and estimated densities.

(b) Calibration transformation.

(c) Normalized Bayes error.

Fig. 2: Estimated densities, calibration transformation and Bayes error plots for the unsupervised SRE 2010 task with PLDA



(a) Target and non-target score histograms, and estimated densities.

(b) Calibration transformation.

(c) Normalized Bayes error.

Fig. 3: Estimated densities, calibration transformation and Bayes error plots for the unsupervised SRE 2010 task with PLDA. Symbol * denotes oracle-initialized models. The score histograms reflect the target trial proportion $\pi = 0.5\%$.

E. Results

To assess the performance of the proposed approach, we start considering the SRE 2019 task with a PLDA classifier that does not employ length normalization. Figure 2-a plots the normalized score histograms, whereas Figure 2-b allows comparing the calibration transformations of the different models with the solution provided by isotonic regression computed through the PAV [43] algorithm on the calibration set. Finally, Figure 2-c shows the corresponding Bayes error plots. C_{llr}, C_{llr}^{ℓ} and C_{prim} are shown in Table I-a. The supervised models have been trained with target weight set to $\zeta = 0.01$.

As expected, the unconstrained NIG model, having more parameters, provides the best fit to the two distributions, and the corresponding non-linear transformation better approximates the PAV curve. The resulting model thus achieves better performance than the linear alternatives. Nevertheless, C-NIG and C-VΓ are also able to provide good estimates of the score distributions, achieving results similar to those of Logistic Regression. Although effective in terms of C_{llr}, in the low FA regions CMLG provides bad calibration even in the supervised task, while C-NIG and C-VΓ deliver very low calibration error.

The benefits of the proposed approaches become more evident in the unsupervised scenarios. Figure 3-a shows the score histograms and the estimated densities for an unsupervised task with a target trial proportion $\pi = 0.5\%$. The corresponding calibration transformation and Bayes error plots are shown in Figures 3-b and 3-c. C_{llr}, C_{llr}^{ℓ} and C_{prim} are shown in Table I-a, together with the results for different target proportions. In this table, and in the following ones, entries marked by a † symbol correspond to performance worse than that of a system based on prior information only.

We can observe that, even with oracle initialization, the CMLG model fails at estimating the distribution of target trials, thus resulting in worse-than-chance performance. In Figure 3-a the CMLG target distribution is not visible, since its mass is concentrated outside of the region that contains the target scores. Indeed, since the non-target right tails decreases faster than the values predicted by CMLG, the model assigns the target scores to the non-target distribution While CMLG fails due to poor modeling assumptions, unconstrained NIG suffers from the opposite issue. The excessive freedom, which was beneficial for the supervised task, allows the model to assign a significant part of non-target scores to the target class. The C-NIG and C-VΓ models, on the other hand, provide good estimates of the target distribution density. From Figure 3-b we can observe that C-VΓ provides the best fit to the PAV curve. It is also worth noting that, in this case, the unconstrained NIG transformation is slightly non-monotonic. The superior performance of the C-VΓ approach is confirmed by the normalized Bayes error plot in Figure 3-c.

Table I-a summarizes the results for different target proportion. The C-NIG results are similar to those we reported in [10]: with oracle-initialization, the model obtains performance equal or better than oracle-initialized NIG for the very low FA regions (C_{prim} metric) for all but the hardest scenario. However, the model tends to provide inaccurate results for low FR regions. Furthermore, the fully unsupervised model is able to reach the performance of the oracle-initialized approach in some, but not all cases. This suggests that better initialization procedures might be necessary for C-NIG. The fully unsupervised C-VΓ model, on the other hand, achieves accuracy similar to that of supervised models in all scenarios. Although for the easier cases $\pi = 0.5\%$ and $\pi = 0.2\%$ the

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

TABLE I: Results for the SRE 2019 task using x-vectors. Oracle-initialized models are denoted by *. A † denotes performance worse than that of a system based only on prior information.

(a) PLDA backend without length normalization.

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.245 — min cost: 0.219				
Supervised	0.263	0.220	0.266	0.253
$\pi = 0.5\%$	†	0.414	0.549 / 0.549	0.288 / 0.289
$\pi = 0.2\%$	†	0.377	0.630 / 0.971	0.264 / 0.264
$\pi = 0.05\%$	†	†	0.970 / 0.880	0.256 / 0.256
C^{ℓ}_{llr}: — Log-Reg: 0.181 — min cost: 0.173				
Supervised	0.212	0.174	0.175	0.183
$\pi = 0.5\%$	†	0.199	0.194 / 0.194	0.183 / 0.183
$\pi = 0.2\%$	†	0.227	0.200 / †	0.190 / 0.190
$\pi = 0.05\%$	†	0.220	0.949 / †	0.199 / 0.199
C_{prim}: — Log-Reg: 0.409 — min cost: 0.405				
Supervised	0.843	0.409	0.406	0.407
$\pi = 0.5\%$	†	0.438	0.405 / 0.405	0.451 / 0.452
$\pi = 0.2\%$	†	0.429	0.405 / †	0.435 / 0.435
$\pi = 0.05\%$	†	0.670	† / 0.438	0.422 / 0.422

(b) PLDA backend with length normalization.

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.205 — min cost: 0.193				
Supervised	0.204	0.195	0.217	0.211
$\pi = 0.5\%$	†	0.299	0.281 / 0.921	0.221 / 0.220
$\pi = 0.2\%$	†	0.580	0.253 / 0.253	0.218 / 0.219
$\pi = 0.05\%$	†	0.674	0.221 / †	0.239 / 0.242
C^{ℓ}_{llr}: — Log-Reg: 0.165 — min cost: 0.161				
Supervised	0.169	0.162	0.163	0.166
$\pi = 0.5\%$	†	0.182	0.164 / 0.865	0.172 / 0.172
$\pi = 0.2\%$	†	0.342	0.163 / 0.163	0.186 / 0.186
$\pi = 0.05\%$	†	0.552	0.173 / †	0.208 / 0.211
C_{prim}: — Log-Reg: 0.418 — min cost: 0.416				
Supervised	0.559	0.417	0.423	0.423
$\pi = 0.5\%$	†	0.442	0.434 / †	0.467 / 0.465
$\pi = 0.2\%$	†	0.422	0.441 / 0.441	0.451 / 0.449
$\pi = 0.05\%$	†	0.589	0.475 / †	0.429 / 0.430

(c) NL-PLDA backend without x-vector scaling

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.212 — min cost: 0.199				
Supervised	0.214	0.201	0.224	0.218
$\pi = 0.5\%$	†	0.345	0.349 / 0.933	0.230 / 0.231
$\pi = 0.2\%$	†	0.550	0.356 / 0.356	0.221 / 0.222
$\pi = 0.05\%$	†	0.692	0.889 / 0.861	0.222 / 0.222
C^{ℓ}_{llr}: — Log-Reg: 0.172 — min cost: 0.166				
Supervised	0.181	0.167	0.169	0.173
$\pi = 0.5\%$	†	0.191	0.175 / 0.886	0.175 / 0.175
$\pi = 0.2\%$	†	0.265	0.175 / 0.175	0.182 / 0.181
$\pi = 0.05\%$	†	0.581	0.813 / †	0.189 / 0.190
C_{prim}: — Log-Reg: 0.400 — min cost: 0.398				
Supervised	0.633	0.400	0.399	0.401
$\pi = 0.5\%$	†	0.421	0.400 / †	0.430 / 0.431
$\pi = 0.2\%$	†	0.402	0.402 / 0.402	0.420 / 0.422
$\pi = 0.05\%$	†	0.691	† / 0.845	0.410 / 0.411

(d) NL-PLDA backend with x-vector scaling

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.194 — min cost: 0.183				
Supervised	0.190	0.184	0.203	0.198
$\pi = 0.5\%$	†	0.190	0.254 / 0.254	0.199 / 0.198
$\pi = 0.2\%$	†	0.191	0.231 / 0.231	0.197 / 0.197
$\pi = 0.05\%$	†	0.315	0.216 / 0.216	0.197 / 0.197
C^{ℓ}_{llr}: — Log-Reg: 0.159 — min cost: 0.156				
Supervised	0.164	0.157	0.158	0.160
$\pi = 0.5\%$	†	0.157	0.160 / 0.160	0.164 / 0.164
$\pi = 0.2\%$	†	0.158	0.158 / 0.158	0.168 / 0.168
$\pi = 0.05\%$	†	0.176	0.158 / 0.158	0.169 / 0.168
C_{prim}: — Log-Reg: 0.392 — min cost: 0.391				
Supervised	0.520	0.392	0.395	0.395
$\pi = 0.5\%$	†	0.392	0.398 / 0.398	0.405 / 0.405
$\pi = 0.2\%$	†	0.392	0.399 / 0.399	0.395 / 0.396
$\pi = 0.05\%$	†	0.393	0.399 / 0.399	0.395 / 0.394

(e) PSVM backend.

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.165 — min cost: 0.156				
Supervised	0.170	0.160	0.169	0.166
$\pi = 0.5\%$	†	0.259	0.216 / 0.919	0.168 / 0.168
$\pi = 0.2\%$	†	0.277	0.216 / 0.216	0.169 / 0.169
$\pi = 0.05\%$	†	0.596	0.211 / 0.847	0.169 / 0.170
C^{ℓ}_{llr}: — Log-Reg: 0.136 — min cost: 0.134				
Supervised	0.147	0.134	0.134	0.136
$\pi = 0.5\%$	†	0.144	0.141 / 0.860	0.139 / 0.139
$\pi = 0.2\%$	†	0.147	0.140 / 0.140	0.142 / 0.142
$\pi = 0.05\%$	†	0.412	0.139 / 0.949	0.142 / 0.142
C_{prim}: — Log-Reg: 0.359 — min cost: 0.352				
Supervised	0.543	0.355	0.354	0.354
$\pi = 0.5\%$	†	0.354	0.354 / †	0.353 / 0.353
$\pi = 0.2\%$	†	0.355	0.354 / 0.354	0.354 / 0.354
$\pi = 0.05\%$	†	0.579	0.353 / 0.403	0.355 / 0.355

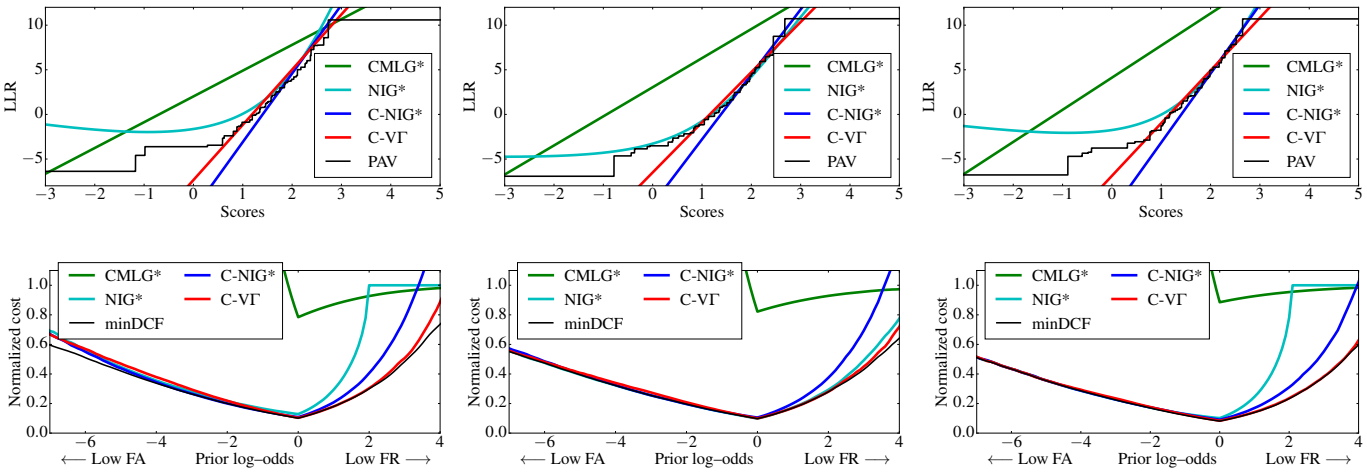
model performs slightly worse than the oracle-initialized NIG or the C-NIG approach in terms of C_{prim}, it significantly outperforms the other models in terms of C^{ℓ}_{llr} and C_{llr}.

Similar results are obtained when length normalization is applied to the embeddings. Even though length normalization modifies slightly the score distributions, the C-VΓ model remains effective. Even without oracle initialization it is able to achieve similar or better performance than the other methods, as can be observed from the results in Table I-b and in Figure 4-a.

The NL-PLDA classifier was introduced to transform speaker vectors so that they better fit the PLDA assumptions. We therefore expected the C-VΓ approach to be well suited for this classifier. This is confirmed by the results in Table I-c, I-d and in Figure 4-b. Since NL-PLDA incorporates an utterance dependent scaling factor similar to length normalization we considered both NL-PLDA without (Table I-c) and with scaling factor (Table I-d and Figure 4-b). In both cases, CMLG fails for all unsupervised scenarios. For scaled NL-

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103



(a) PLDA with length normalization. (b) NL-PLDA with scaling factor. (c) PSVM.

Fig. 4: Calibration transformations and normalized Bayes error plot for the unsupervised SRE 2019 task using x-vectors. Symbol * denotes oracle-initialized models. The target proportion is $\pi = 0.5\%$.

PLDA (Table I-d) the C-NIG results are similar to the PLDA ones. The model provides very good calibration in the low FA region, but is less accurate for low FR. Unconstrained NIG is more effective for the easier tasks, but significantly worse for the hardest $\pi = 0.05\%$ scenario. Without x-vector scaling (Table I-c), the two models have similar behavior, and provide good calibration only for the easier tasks and in the low FA region. C-VΓ proves more robust, achieving close to optimal calibration in both cases even without oracle-initialization.

Finally, the results for the PSVM classifier are shown in Table I-e and Figure 4-c. As for PLDA with length normalization, oracle-initialized C-NIG is more effective than oracle-initialized NIG. Both provide very good results for low FA. On the other hand, C-NIG performs better than NIG, but both are not accurate, in the low FR region. Moreover, also in this case we were not able to achieve the same accuracy with fully unsupervised C-NIG models for some target proportions. On the contrary, fully unsupervised C-VΓ provides very good calibration for a wide range of operating points and all considered tasks.

A second set of experiments on the SRE 2010 and SRE 2012 tasks confirm the robustness of the proposed approach. Figure 5-a plots the score distribution and the estimated densities for the different models for the SRE 2010 task with i-vectors and a PLDA backend. In this case all models approximately locate the target scores. However, the CMLG model overestimates the non-target distribution right tail, whereas the unconstrained NIG approach significantly under-estimates the left tail of the target distribution. The C-NIG and C-VΓ both provide a better fit of the two histograms, and the latter more accurately models the distribution tails. As can be observed in Figures 5-b and 5-c and in Table II, the CMLG transformation leads to good C_{llr}, but results in a significant degradation in terms of C_{prim}, whereas the unconstrained NIG model provides accurate calibration only for the very low FA regions. The C-NIG and C-VΓ models are both very accurate in terms of C_{prim}, but the latter is much more effective for a large range of operating points and results in significantly lower C_{llr}.

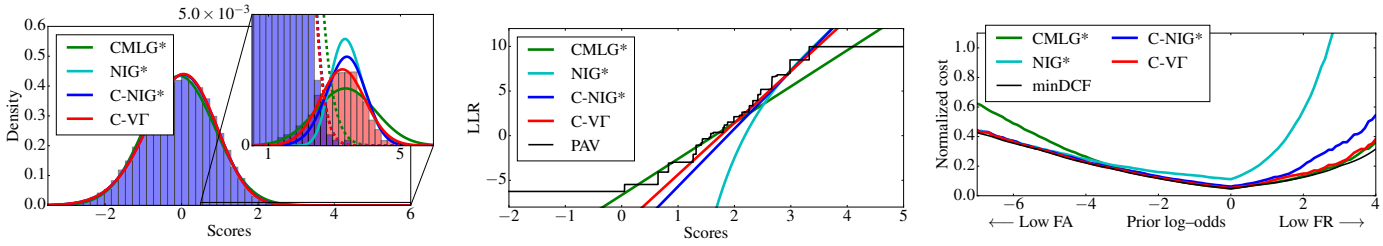
TABLE II: Results for the SRE 2010 task using i-vectors and PLDA. Oracle-initialized models are denoted by *. A † denotes performance worse than that of a system based only on prior information.

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.098 — min cost: 0.094				
Supervised	0.103	0.097	0.101	0.100
$\pi = 0.5\%$	0.100	0.715	0.133 / 0.134	0.113 / 0.108
$\pi = 0.2\%$	0.102	†	0.150 / 0.148	0.117 / 0.111
C^{ℓ}_{llr}: — Log-Reg: 0.090 — min cost: 0.089				
Supervised	0.094	0.091	0.091	0.091
$\pi = 0.5\%$	0.096	0.144	0.103 / 0.104	0.096 / 0.094
$\pi = 0.2\%$	0.100	0.177	0.109 / 0.108	0.098 / 0.095
C_{prim}: — Log-Reg: 0.443 — min cost: 0.420				
Supervised	0.597	0.441	0.428	0.437
$\pi = 0.5\%$	0.614	0.440	0.431 / 0.432	0.432 / 0.428
$\pi = 0.2\%$	0.612	0.440	0.428 / 0.429	0.443 / 0.439

The results for SRE 2012 reported in Figure 6 and Table III show a similar trend. Supervised models achieve similar C_{llr}, but the CMLG approach suffers in the low FA regions, obtaining significantly worse C_{prim}. In the unsupervised scenarios with low target proportions CMLG is competitive in terms of C_{llr} for NL-PLDA backend, but provides bad results for the PSVM backend. Furthermore, the calibration error in the low FA regions is significant in all cases. For the oracle-initialized, unconstrained NIG approach, we also observe inconsistent C_{llr} results: the model is accurate for PSVM scores, but provides the worst results for NL-PLDA. The oracle-initialized C-NIG is much more effective. In this case the fully initialized models achieve the same accuracy as the oracle-initialized ones. Once again, the fully unsupervised C-VΓ model outperforms C-NIG and achieves close to optimal

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103



(a) Normalized target and non-target score histograms, and estimated densities.

(b) Calibration transformation.

(c) Normalized Bayes error.

Fig. 5: Estimated densities, calibration transformation and Bayes error plots for the unsupervised SRE 2010 task with PLDA. Symbol * denotes oracle-initialized models. The target proportion is $\pi = 0.5\%$.

TABLE III: Results for the SRE 2012 task using e-vectors. Oracle-initialized models are denoted by *. A † denotes performance worse than that of a system based only on prior information.

(a) PSVM

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.061 — min cost: 0.057				
Supervised	0.065	0.062	0.062	0.062
$\pi = 0.5\%$	0.080	0.067	0.089 / 0.089	0.076 / 0.078
$\pi = 0.2\%$	†	0.074	0.117 / 0.117	0.087 / 0.084
C^{ℓ}_{llr}: — Log-Reg: 0.058 — min cost: 0.057				
Supervised	0.063	0.058	0.059	0.058
$\pi = 0.5\%$	0.080	0.059	0.069 / 0.069	0.063 / 0.063
$\pi = 0.2\%$	†	0.061	0.078 / 0.078	0.066 / 0.065
C_{prim}: — Log-Reg: 0.212 — min cost: 0.210				
Supervised	0.317	0.212	0.217	0.214
$\pi = 0.5\%$	0.347	0.211	0.224 / 0.224	0.215 / 0.215
$\pi = 0.2\%$	†	0.212	0.228 / 0.228	0.217 / 0.216

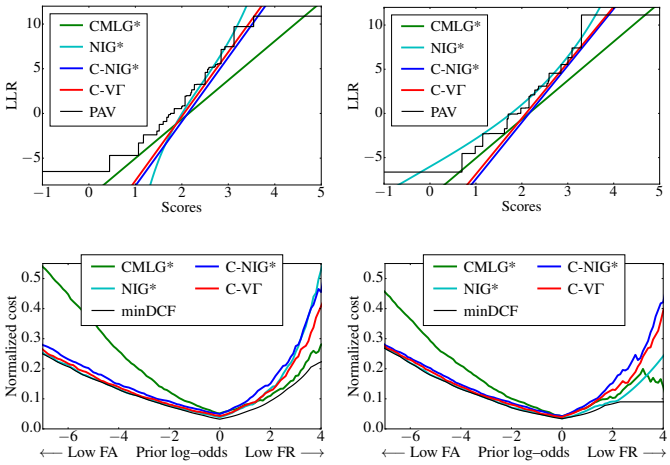
(b) NL-PLDA

	CMLG*	NIG*	C-NIG* / C-NIG	C-VΓ* / C-VΓ
C_{llr}: — Log-Reg: 0.068 — min cost: 0.064				
Supervised	0.073	0.067	0.069	0.068
$\pi = 0.5\%$	0.095	0.130	0.106 / 0.106	0.089 / 0.091
$\pi = 0.2\%$	0.114	0.148	0.157 / 0.157	0.115 / 0.111
C^{ℓ}_{llr}: — Log-Reg: 0.059 — min cost: 0.058				
Supervised	0.068	0.059	0.060	0.059
$\pi = 0.5\%$	0.096	0.063	0.074 / 0.074	0.066 / 0.067
$\pi = 0.2\%$	0.109	0.062	0.087 / 0.087	0.073 / 0.072
C_{prim}: — Log-Reg: 0.203 — min cost: 0.199				
Supervised	0.376	0.200	0.214	0.208
$\pi = 0.5\%$	0.423	0.201	0.226 / 0.226	0.208 / 0.208
$\pi = 0.2\%$	0.434	0.200	0.230 / 0.230	0.207 / 0.206

calibration in the low FA regions.

VII. Conclusions

We analyzed the theoretical distribution of the scores of a PLDA classifier, showing that Generalized Hyperbolic, and in particular Variance-Gamma distributions, are good candidates for modeling scores generated by PLDA-like models. We introduced the Constrained GH calibration approach, that includes as special cases our previous C-NIG and the novel C-VΓ model. The experimental results show that these methods, in particular the latter, are able to provide good calibration accuracy even for unsupervised scenarios with unbalanced target proportions. C-VΓ outperforms not only CMLG, but also our previous C-NIG method. In the future we plan to refine these analyses focusing on the effects of train and test mismatch on the score distributions. This would allow us to incorporate, for example, duration or noise effects at calibration level, reducing not only actual but also minimum detection costs. Furthermore, we believe that accurate models of score distributions may have the potential for improving score normalization approaches.



(a) NL-PLDA

(b) PSVM

Fig. 6: Calibration transformations and Bayes error plots for the unsupervised SRE 2012 task. Symbol * denotes oracle-initialized models. The target proportion is $\pi = 0.5\%$.

VIII. Acknowledgments

I would like to thank Prof. Pietro Laface for useful discussions and preliminary review of this work. Computational resources were provided by HPC@POLITO (http://www.hpc.polito.it)

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2020.3040103

REFERENCES

[1] N. Brummer *et al.*, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *Transactions on Audio, Speech and Language Processing*, 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 *et al.*, "Score calibration in face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 246–256, 2014.

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

[8] 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.

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

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

[11] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," in *Proceedings of ICASSP 2014*, pp. 1680–1684, 2014.

[12] 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.

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

[14] O. E. Barndorff-Nielsen, "Exponentially decreasing distributions for the logarithm of particle size," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 353, pp. 401 – 419, 1977.

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

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

[17] "The NIST year 2010 speaker recognition evaluation plan," 2010. Available at http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.

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

[19] 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.

[20] D. Snyder *et al.*, "X–vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of ICASSP 2018*, pp. 5329–5333, 2018.

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

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

[23] S. Cumani *et al.*, "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] K. Slooten and R. Meester, "Forensic identification: Database likelihood ratios and familial DNA searching," in *arXiv:1201.4261 [stat.AP]*, 2012.

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

[27] O. E. Barndorff-Nielsen, "Normal inverse Gaussian distributions and stochastic volatility modelling," *Scandinavian Journal of Statistics*, vol. 24, no. 1, pp. 1–13, 1997.

[28] M. S. Paolella, *Intermediate Probability: A Computational Approach*. Wiley-Interscience, 2007.

[29] P. Blæsild, "The two-dimensional hyperbolic distribution and related distributions, with an application to Johannsen's bean data," *Biometrika*, vol. 68, no. 1, pp. 251–263, 1981.

[30] C. G. Broyden, "The convergence of a class of double rank minimization algorithms: 2. the new algorithm," *IMA Journal of Applied Mathematics*, vol. 6, no. 3, pp. 222–231, 1970.

[31] R. Fletcher, "A new approach to variable metric algorithms," *Computer Journal*, vol. 13, no. 3, pp. 317–322, 1970.

[32] D. Goldfarb, "A family of variable metric methods derived by variational means," *Mathematics of Computation*, vol. 24, no. 109, pp. 23–26, 1970.

[33] D. F. Shanno, "Conditioning of quasi-newton methods for function minimization," *Mathematics of Computation*, vol. 24, no. 11, pp. 647–650, 1970.

[34] B. Jrgensen, *Statistical properties of the generalized inverse Gaussian distribution*. No. 9 in Lecture notes in statistics, Springer, 1982.

[35] M. Jamshidian and R. I. Jennrich, "Acceleration of the EM algorithm by using Quasi-Newton methods," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 59, no. 3, pp. 569–587, 1997.

[36] D. Snyder *et al.*, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proceedings of ICASSP 2019*, pp. 5796–5800, May 2019.

[37] 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.

[38] P. Kenny *et al.*, "Deep Neural Networks for extracting Baum-Welch statistics for speaker recognition," in *Proceedings of Odyssey 2014*, pp. 293–298, 2014.

[39] D. Garcia-Romero and A. McCree, "Insights into Deep Neural Networks for speaker recognition," in *Proceedings of Interspeech 2015*, pp. 1141–1145, 2015.

[40] 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.

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

[42] 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.

[43] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 02, (New York, NY, USA), p. 694699, Association for Computing Machinery, 2002.

Copyright (c) 2021 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.