INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Normal variance–mean mixtures for unsupervised score calibration

Sandro Cumani

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

Abstract

Generative calibration models have shown to be an effective alternative to traditional discriminative score calibration techniques, such as Logistic Regression (LogReg). Provided that the score distribution assumptions are sufficiently accurate, generative approaches not only have similar or better performance with respect to LogReg, but also allow for unsupervised or semi–supervised training.

Recently, we have proposed non–Gaussian linear calibration models able to overcome the limitations of Gaussian approaches. Although these models allow for better characterization of score distributions, they still require the target and non–target distributions to be reciprocally symmetric.

In this work we further extend these models to cover asymmetric score distributions, as to improve calibration for both supervised and unsupervised scenarios. The improvements have been assessed on NIST SRE 2010 telephone data.

Index Terms: score calibration, likelihood ratio, linear score calibration models, unsupervised training

1. Introduction

The output of a speaker verification system is a score that allows discriminating between the same speaker (target trial, higher score values) and different speaker (non–target trial, lower score values) hypotheses. Given an application, hard decisions require comparing the score with a suitable threshold. For systems whose output can be interpreted as a log–likelihood ratio (LLR) between the two hypotheses, the optimal threshold depends only on prior probabilities, and the costs associated to the different classification errors. In practice, often the outputs of a recognizer cannot be interpreted as log–likelihood ratios. This may be due to the intrinsic nature of the classifier (e.g. Support Vector Machines [1, 2]). However, even for statistical models such as Probabilistic Linear Discriminant Analysis (PLDA) [3, 4], mismatches between the training and evaluation populations or imprecise model assumptions can impair the LLR interpretation of scores. For these reasons, score calibration techniques are employed to transform the scores produced by a recognizer so that they can be interpreted as well–calibrated LLRs.

The standard approach for score calibration is based on discriminative prior–weighted Logistic Regression (LogReg) [5], which optimizes the expected value of the logarithmic proper scoring rule [6] assuming a linear calibration model. This technique has been successfully applied to different tasks in the past [7–10]. LogReg relies on a labelled dataset that closely matches the testing conditions.

Recently, alternative approaches based on generative models have gained interest [11–13]. Although these methods achieve results similar to LogReg in supervised scenarios, they provide more insights on the behaviour of well–calibrated scores and can be easily extended to handle missing labels [14].

In [11] the authors have analyzed the theoretical properties of log–likelihood ratios, showing that well–calibrated LLRs satisfy the "LLR of the LLR is the LLR" condition. The LLR property constrains the admissible distributions that can generate well–calibrated scores. The Constrained Maximum Likelihood Gaussian (CMLG) approach was proposed in [11], based on the assumption of a linear calibration model and that target and non–target scores are samples of two Gaussian distributions, with tied parameters to satisfy the LLR constraint. CMLG was extended to handle missing labels in [14] by considering a two–component Gaussian Mixture Model (GMM), whose parameters are tied as in CMLG.

The Gaussian assumption, however, is often inaccurate. This can result in a degradation of the performance of CMLG, and the impact can be even more relevant for unsupervised training. Indeed, unsupervised scenarios are usually characterized by very unbalanced trial distributions, as non–target training scores are the vast majority. Target scores can, thus, be confused with non–target scores generated by the non–target distribution tails. In this case, proper modelling of the tail behaviour of the score distributions becomes essential.

Non–Gaussian generative models have been recently proposed for both linear and non–linear calibration. The work [12] analyzes models based on different distributions, including T–student and Normal Inverse Gaussians (NIG) [15]. Rather than assuming a specific calibration model as in [11], these approaches estimate a probabilistic model in score space, and compute the LLRs by evaluating the likelihood ratio between the hypotheses that a score was generated by the target or by the non–target distribution, respectively. This results in non–linear, and possibly non monotonic, score mappings.

In [13], we have proposed an alternative non–Gaussian approach that extends CMLG showing that normal variance–mean mixtures, which include NIG as a particular case, are suited to model well–calibrated LLRs, provided that the distribution parameters are tied as required by the LLR constraint. In this work, we extend these results to unsupervised training, following the same approach of [14], but relying on the more powerful class of variance–mean mixture distributions. We start by addressing an important limitation of our previous approach, namely the assumption that target and non–target distributions are reciprocally symmetric. We show that symmetry is not required to satisfy the LLR constraint, and that a broader class of variance–mean mixture distributions can be employed to represent well–calibrated LLRs. We then extend the method to handle missing labels. We show that, as expected, this leads to a significant improvement with respect to unsupervised CMLG for non–Gaussian distributed scores. We also contrast our method with an unsupervised extension of the NIG model of [12], and comment on the results obtained by the two techniques.

The rest of the paper is organized as follows. Section 2 recalls the CMLG model. Section 3 extends our tied NIG ap-

http://dx.doi.org/10.21437/Interspeech.2019-1609

proach to model asymmetric distributions and to handle missing labels. Experimental results are illustrated in Section 4, and conclusions are given in Section 5.

2. CMLG Calibration

Generative score models interpret scores as samples of a random variable X, whose conditional distributions given the target (T) and the non–target (F) classes are $f_{X|T}$ and $f_{X|F}$, respectively. In [11, 16] it was shown that, if the scores are well–calibrated LLRs, they satisfy the LLR constraint

$$e^x = \frac{f_{X|T}(x)}{f_{X|F}(x)} \tag{1}$$

The CMLG approach [11] assumes that the observed scores s are obtained by an affine transformation of Gaussian–distributed well–calibrated scores $x \sim X$. The parameters of the distributions of $X|T$ and $X|F$ are tied to satisfy (1):

$$f_{X|T}(x) = \mathcal{N}(x|\mu, 2\mu)\ , \quad f_{X|F}(x) = \mathcal{N}(x|-\mu, 2\mu)\ , \tag{2}$$

where μ is the mean of the target distribution. A weighted Maximum Likelihood (ML) criterion can be used to estimate both the model parameter μ and the affine score transformation.

The approach was extended in [14] to handle missing training labels. Supervised CMLG defines the conditional distributions for $X|T$ and $X|F$. Assuming that prior distributions for target class is w_T, the density f_X of X is given by the two–components GMM:

$$f_X(x) = w_T\mathcal{N}(x|\mu, 2\mu) + (1-w_T)\mathcal{N}(x|-\mu, 2\mu)\ . \tag{3}$$

Assuming linear calibration, the distribution of observed scores is again a GMM. The EM algorithm can then be used to estimate the calibration parameters, the parameter μ and the target class proportion w_T.

3. Variance–mean mixtures for calibration

In many cases the Gaussian assumptions of CMLG are not realistic, as score distributions often are skewed, have significantly different variances, and show non–Gaussian tail behaviours. In these scenarios CMLG does not provide very accurate results, especially when score labels are not available for training, as we will show in the experimental section.

3.1. Symmetric variance–mean mixtures

In [13] we have shown that variance–mean mixtures are good candidates for modelling the distributions of well–calibrated LLRs, and we gave an interpretation of the corresponding model in terms of a generative procedure for sampling the target and non–target scores. The model considers variance–mean mixture distributions for target and non–target scores:

$$\begin{aligned} X|T &= \mu_T + \beta_T(V_T) + \sqrt{(V_T)}Y\ , \\ X|F &= \mu_F + \beta_F(V_F) + \sqrt{(V_F)}Y\ , \end{aligned} \tag{4}$$

where V_T, V_F and Y are independent random variables, Y has a standard normal distribution, and the densities of V_T and V_F are $g_{V_T}(v)$ and $g_{V_F}(v)$, respectively. We showed that sufficient conditions for satisfying the LLR constraint are:

$$g_{V_T} = g_{V_F}\ ,\ \mu_T = \mu_F = 0\ ,\ \beta_T = \frac{1}{2}\ ,\ \beta_F = -\frac{1}{2}\ . \tag{5}$$

This implies that target and non–target densities are reciprocally symmetric, i.e.

$$f_{X|F}(x) = f_{X|T}(-x) \tag{6}$$

3.2. Asymmetric variance–mean mixtures

In the following we show that the model of [13] can be extended to cover a wider range of distributions, allowing for better score models. As in our previous work, we focus on the class of Generalized Hyperbolic (GH) distributions

$$\begin{aligned} GH(x|\lambda,\alpha,\beta,\delta,\mu) = Z(\lambda,\alpha,\beta,\delta)\left[\delta^2+(x-\mu)^2\right]^{\frac{\lambda-\frac{1}{2}}{2}} \\ \cdot\, e^{\beta(x-\mu)}K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)\ , \end{aligned} \tag{7}$$

where K_ν denotes the modified Bessel function of the third kind of order ν and Z a normalization constant. In particular, we consider the NIG subclass of GH distributions[1], obtained by setting $\lambda = -\frac{1}{2}$. We thus assume that

$$f_{X|F}(x) = NIG(x|\alpha,\beta,\delta,\mu) = GH(x|-\frac{1}{2},\alpha,\beta,\delta,\mu)\ . \tag{8}$$

From the LLR constraint (1), the Moment Generating Functions (m.g.f.) of $X|T$ and $X|F$ are related by:

$$\begin{aligned} M_{X|T}(z) &\triangleq \int e^{zx} f_{X|T}(x)dx = \int e^{(z+1)x} f_{X|F}(x)dx \\ &= M_{X|F}(z+1)\ . \end{aligned} \tag{9}$$

Since $f_{X|T}$ must be a proper density function, the LLR constraint requires that $M_{X|F}$ satisfies

$$M_{X|F}(1) = M_{X|T}(0) = \int f_{X|T}(x)dx = 1\ . \tag{10}$$

The m.g.f. of $X|F$ is given by [15]:

$$M_{X|F}(z) = e^{\mu z + \delta\sqrt{\alpha^2-\beta^2} - \delta\sqrt{\alpha^2-(\beta+z)^2}}\ . \tag{11}$$

Equation (10) requires thus that $\alpha^2 \geq (\beta+1)^2$ and

$$\mu = \delta\sqrt{\alpha^2-(\beta+1)^2} - \delta\sqrt{\alpha^2-\beta^2} = \delta(\gamma_T - \gamma_F)\ , \tag{12}$$

with $\gamma_F = \sqrt{\alpha^2-\beta^2}$ and $\gamma_T = \sqrt{\alpha^2-(\beta+1)^2}$. The m.g.f. of $X|T$ can be computed from (9) and (11):

$$M_{X|T}(z) = M_{X|F}(z+1) = e^{\mu+\mu z+\delta\gamma_F - \delta\sqrt{\alpha^2-(\beta+z+1)^2}} \tag{13}$$

and, since from (12) $\mu + \delta\gamma_F = \delta\gamma_T = \delta\sqrt{\alpha^2-(\beta+1)^2}$:

$$M_{X|T}(z) = e^{\mu z + \delta\sqrt{\alpha^2-(\beta+1)^2} - \delta\sqrt{\alpha^2-[(\beta+1)+z]^2}}\ . \tag{14}$$

Assuming $\alpha^2 > (\beta+1)^2$ the m.g.f. of (14) uniquely identifies the NIG density[2]:

$$f_{X|T}(x) = NIG(x|\alpha,\beta+1,\delta,\mu)\ . \tag{15}$$

Well–calibrated LLRs can be modelled by two NIG distributions whose parameters are tied:

$$\begin{aligned} X|F &\sim NIG\Big(\alpha,\beta,\delta,\delta\left(\gamma_T-\gamma_F\right)\Big)\ , \\ X|T &\sim NIG\Big(\alpha,\beta+1,\delta,\delta\left(\gamma_T-\gamma_F\right)\Big)\ . \end{aligned} \tag{16}$$

[1] Similar consideration can be derived for the general case

[2] We require $\alpha^2 > (\beta+1)^2$ rather than $\alpha^2 \geq (\beta+1)^2$ so that the m.g.f. $M_{X|T}$ is defined in an open interval around 0, and therefore uniquely identifies the density of $X|T$

It is worth noting that setting $\beta = -\frac{1}{2}$ results in $\gamma_T = \gamma_F$, implying $\mu = 0$, which recovers the model of [13]. However, model (16) does not require the value of β to be fixed, and allows characterizing non–specular distributions.

Assuming a linear calibration model $x(s) = as + b$, where $a > 0$, and $x(s)$ denotes the calibrated score, the random variables that generated the observed scores are

$$S|T = \frac{1}{a}(X|T) - \frac{b}{a}\,, \quad S|F = \frac{1}{a}(X|F) - \frac{b}{a}\,, \tag{17}$$

with distributions given by [17]:

$$\begin{aligned} S|F \sim NIG(\overline{\alpha}, \overline{\beta}_F, \overline{\delta}, \overline{\mu})\,, \;\; S|T \sim NIG(\overline{\alpha}, \overline{\beta}_T, \overline{\delta}, \overline{\mu})\,, \\ \overline{\alpha} = a\alpha\,, \;\; \overline{\beta}_F = a\beta\,, \;\; \overline{\beta}_T = a\beta + a\,, \;\; \overline{\delta} = a\delta \\ \overline{\mu} = \frac{\delta(\gamma_T - \gamma_F) - b}{a}\,. \end{aligned} \tag{18}$$

The parameters $(\overline{\alpha}, \overline{\beta}_F, \overline{\beta}_T, \overline{\delta}, \overline{\mu})$ can be estimate by maximizing the (ρ–weighted) log–likelihood

$$\frac{\rho}{N_T}\sum_{i=1}^{N_T}\log f_{S|T}(s_{T,i}) + \frac{1-\rho}{N_F}\sum_{i=1}^{N_F}\log f_{S|F}(s_{F,i}), \tag{19}$$

where $s_{T,i}$ and $s_{F,i}$ denote target and non–target scores, respectively. The ML solution can be obtained by the Expectation Maximization algorithm [18] for NIG distributions, modified to account for the shared parameters $(\overline{\alpha}, \overline{\delta}, \overline{\mu})$. Given $(\overline{\alpha}, \overline{\beta}_F, \overline{\beta}_T, \overline{\delta}, \overline{\mu})$ we can obtain the calibration parameters as:

$$\begin{aligned} a = \overline{\beta}_T - \overline{\beta}_F\,, \\ b = \overline{\delta}\left(\sqrt{\overline{\alpha}^2 - \overline{\beta}_T^2} - \sqrt{\overline{\alpha}^2 - \overline{\beta}_F^2}\right) - \left(\overline{\beta}_T - \overline{\beta}_F\right)\overline{\mu}\,. \end{aligned} \tag{20}$$

3.3. Unsupervised variance–mean mixture calibration

Model (18) can be extended to handle missing labels following the same strategy of [14]. Assuming that the prior probability for the target class is w_T, the scores can be interpreted as samples of random variable S, with p.d.f. given by:

$$f_S(s) = w_T NIG(s|\overline{\alpha}, \overline{\beta}_T, \overline{\delta}, \overline{\mu}) + w_F NIG(s|\overline{\alpha}, \overline{\beta}_F, \overline{\delta}, \overline{\mu})\,, \tag{21}$$

i.e. a two–component mixture model whose components are the NIG densities of $S|F$ and $S|T$. The model parameters can be estimated by maximizing the log–likelihood:

$$\mathscr{L}(\overline{\alpha}, \overline{\beta}_F, \overline{\beta}_T, \overline{\delta}, \overline{\mu}) = \sum_i \log f_S(s_i) \tag{22}$$

using the EM algorithm, as in the Gaussian case.

3.4. Unsupervised unconstrained NIG calibration

An unconstrained NIG model for supervised calibration was presented in [12]. The model is estimated by independently fitting two unconstrained NIG distributions to target and non–target scores, respectively. The calibration transformation is then obtained by computing the log–likelihood ratio between the two densities. This results in non–linear calibration models that provide good accuracy for supervised scenarios, though the corresponding calibration transformation is not necessarily monotonic. The unsupervised extension that we consider is based on a mixture of two NIG distributions

$$\begin{aligned} f_S(s) = w_T NIG(s|\alpha_T, \beta_T, \delta_T, \mu_T) \\ + w_F NIG(s|\alpha_F, \beta_F, \delta_F, \mu_F)\,. \end{aligned} \tag{23}$$

The model is similar to (21), but parameters are not tied.

4. Results

In this section we compare the performance of CMLG, unconstrained NIG and our Constrained Maximum Likelihood NIG (CMLNIG) method for both supervised and unsupervised scenarios. We refer to [13] for comparisons with LogReg models for supervised tasks.

Since we are interested in non–Gaussian distributed scores, we focus on variable duration utterances, and we analyze the scores of an i–vector [19] based PLDA model that exploits the i–vector uncertainty [20, 21].

The system is based on 400–dimensional i–vectors extracted from a gender independent, 2048 components, GMM with diagonal covariances. The training set for the UBM consisted of NIST 04, 05 and 06 data. Switchboard 2 was added for training the i–vector extractor. The tests were performed on the female portion of SRE 2010 tel–tel extended condition, cutting short segments from 3 to 60 seconds.

All EM–based models have been trained using a Quasi–Newton accelerator for the EM algorithm [22].

4.1. Supervised calibration

In this section we compare the performance of the generative approaches for a supervised scenario. The model parameters have been estimated on a subset of the NIST 08 female dataset, whose segments were cut to match the durations of the test set.

The results are shown in Figure 1 in terms of normalized Bayes error rate [23]. X–axis corresponds to different target prior log–odds $x = \log\frac{p}{1-p}$, where p is a synthetic prior. Y–axis plots the corresponding normalized actual DCF. The likelihood weighting parameter ρ was set to 0.5 for the left plot, and to 0.1 for the right plot. Labels "CMLNIG–S" and "CMLNIG–A" refer to the symmetric (Section 3.1) and asymmetric (Section 3.2) versions of our model. Label "NIG" refers to the unconstrained NIG model.

The results show that all models, with the exception of CMLG with $\rho = 0.1$, achieve good calibration for a wide range of operating points, with minimal differences among the three NIG–based methods. Furthermore, the choice of ρ does not significantly affect CMLNIG methods.

4.2. Unsupervised calibration

For the unsupervised scenario, we randomly selected 25% of the evaluation scores for training. This mimics real applications where unlabelled but matching data are available.

Since likelihood optimization can get stuck in bad local optima, a first set of tests were performed initializing the models with the parameters obtained by supervised training on SRE08 (oracle initialization in the following). The results are shown in Figure 2. Different target proportions were simulated by artificially re–weighting the non–target scores. The central plot

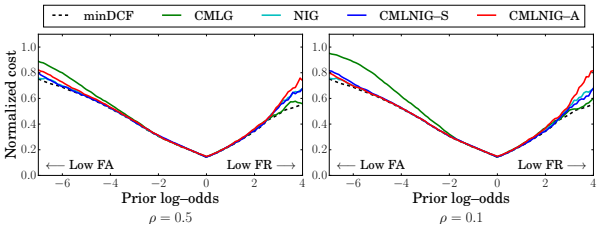

Figure 1: *Normalized Bayes errors — supervised training.*

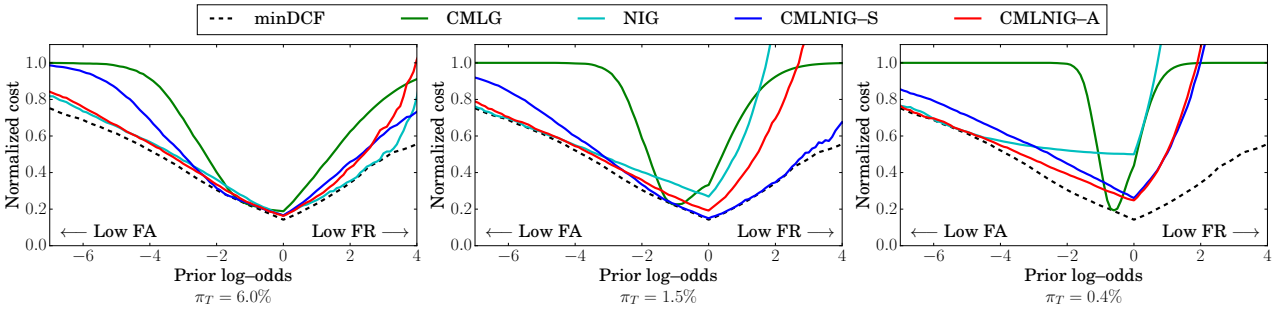

Figure 2: *Normalized Bayes error — oracle–initialized unsupervised training. π_T denotes the proportion of training targets.*

corresponds to the original target proportion $\pi_T = 1.5\%$. Left and right plots refer to the simulated proportions $\pi_T = 6.0\%$ and $\pi_T = 0.4\%$, respectively.

The first observation is that CMLG is not able to provide good calibration even for the easiest $\pi_T = 6.0\%$ scenario. This is expected, as CMLG cannot properly model the non–Gaussian score distributions. The symmetric CMLNIG–S model improves calibration with respect to CMLG, but is still inaccurate for both very low False Alarms (FA) and low False Rejections (FR). The assumption that the target and non–target distributions are reciprocally symmetric is inaccurate for this training set, thus the model is not able to correctly identify the two components. On the contrary, both NIG and CMLNIG–A models achieve good calibration over a wide range of operating points.

As the proportion of target trials reduces, the performance of the different methods degrades. For $\pi_T = 1.5\%$ both NIG and CMLNIG–A still achieve good calibration for the very low FA region, with CMLNIG–A achieving lower actual DCFs for a larger set of operating points. However, both models incur in a severe degradation for the low FR region. Surprisingly, the CMLNIG–S model is able to achieve better results for low FR, but provides worse calibration for low FA. For the hardest scenario, $\pi_T = 0.4\%$, unsupervised training does not allow recovering well-calibrated scores. Nevertheless, CMLNIG–A still achieves better performance than other NIG–based approaches.

A second set of experiments was performed to analyze unsupervised training of NIG–based models without oracle initializations. Due to the difficulty of obtaining good initial estimates of the parameters, the following strategy was adopted. The initial models were obtained by splitting a single Gaussian estimated on the pooled scores. We then proceeded by estimating the parameters of the CMLNIG–S model, which were then used for the CMLNIG–A model initialization. Finally, the NIG model was initialized from the CMLNIG–A parameters. The results, illustrated in Figure 3, show that the CMLNIG models converge to the same local optima we obtained when using the supervised initialization. On the contrary, the NIG model (solid cyan line) is more sensitive to the selection of proper initial values. Indeed, for the $\pi_T = 1.5\%$ we were not able to obtain the same calibration accuracy of the model based on supervised initialization (dashed cyan line).

We would like to highlight that different training data and target proportions can lead to slightly different results, and in some cases the aforementioned training strategy allows unconstrained NIG models to achieve similar results as CMLNIG–A. However, the unconstrained NIG model is more prone to converge to solutions that do not properly characterize the target and non–target score distributions. This is due a larger amount of parameters and a more complex calibration transformation. On the contrary, the proposed approach, CMLNIG–A, combines powerful distributions with a simpler, linear calibration model. This allows improving the estimation of the distribution of small amounts of target trials.

5. Conclusions

We proposed a linear calibration approach for unsupervised calibration in presence of skewed and asymmetric score distributions. The scores are modelled as samples of two NIG distributions, whose parameters are tied to satisfy the LLR constraint.

The proposed approach is able to outperform both the CMLG and unconstrained NIG methods, and, as long as the target proportion in the training set is sufficiently large, produces well–calibrated scores for a large set of operating points.

Nevertheless, our results show that unsupervised calibration is a challenging task, and requires further investigation on the score distributions of verification systems.

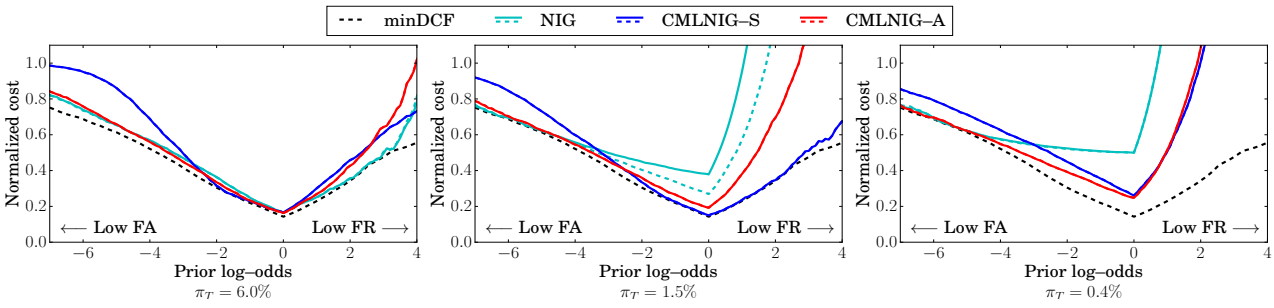

Figure 3: *Normalized Bayes error — unsupervised training. Solid lines refer to fully unsupervised training, dashed lines refer to oracle–initialized unsupervised training. π_T denotes the proportion of training targets.*

6. References

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

[2] 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.

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

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

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

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

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

[8] 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.

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

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

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

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

[13] S. Cumani and P. Laface, "Tied normal variance–mean mixtures for linear score calibration," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6121–6125.

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

[15] O. E. Barndorff-Nielsen, "Normal inverse Gaussian distributions and stochastic volatility modelling," *Scandinavian Journal of Statistics*, vol. 24, no. 1, pp. 1–13, 1997. [Online]. Available: http://www.jstor.org/stable/4616433

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

[17] 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. [Online]. Available: http://www.jstor.org/stable/2335826

[18] D. Karlis, "An EM type algorithm for maximum likelihood estimation of the normal–inverse Gaussian distribution," *Statistics & Probability Letters*, vol. 57, no. 1, pp. 43–52, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167715202000408

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

[20] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in Probabilistic Linear Discriminant Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 846–857, 2014.

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

[22] 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. [Online]. Available: http://www.jstor.org/stable/2346010

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