Impostor Score Statistics as Quality Measures for the Calibration of Speaker Verification Systems

The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

Impostor score statistics as quality measures for the calibration of speaker verification systems

Sandro Cumani, Salvatore Sarni

Department of Control and Computer Engineering
Politecnico di Torino, Italy
sandro.cumani@polito.it, salvatore.sarni@polito.it

Abstract

Trial-dependent miscalibration can severely affect the performance of speaker verification systems. Global calibration methods address the problem by incorporating side-information into the calibration model. Alternatively, score normalization approaches exploit statistics computed from scores of impostor cohorts. While effective in some scenarios, the latter approaches suffer from poor global calibration, and in some cases may even increase trial-dependent miscalibration with respect to unnormalized scores. While the former issue can be addressed through global calibration, the latter problem can result in degraded performance. In this work, we provide a theoretical framework for incorporating impostor score statistics as side information in discriminative calibration models. Our approach allows us to improve both global and trial-dependent calibration, without incurring in some of the issues of score normalization. Results on SRE 2019 and SITW datasets show that our approach achieves similar or better (up to 15% relative) results compared to state-of-the-art score normalization techniques. The model can also be trivially extended to incorporate additional side-information.

1. Introduction

Calibrated speaker verification systems classify a trial as belonging to a single speaker (same speaker or target trial) or to different speakers (non-target trial) by computing a score that represents the log-likelihood ratio (LLR) between the target and non-target hypotheses. The score is then compared with a threshold that depends on class priors and misclassification costs to produce a hard decision. Mismatch between the training and the evaluation populations, or the intrinsic properties of the classification back-end, however, may result in verification scores that are not well calibrated. In this case, hard decisions based only on class priors and error costs are sub-optimal, and can result in significantly larger misclassification costs. Several approaches have been proposed to address miscalibration. Discriminative methods based on prior-weighted Logistic Regression (LogReg) [1, 2] are often employed, and good results are also obtained by generative supervised and unsupervised techniques based on Gaussian assumptions [3, 4] or more complex distributions [5, 6, 7, 8]. For an in-depth analysis of generative calibration we refer the reader to [6]. Both LogReg and the aforementioned generative approaches provide global, or dataset-level, calibration: scores are mapped to LLRs through a transformation that depends only on the score itself, and is often assumed to be monotone as to avoid modifying the discrimination power of the original scores. Global calibration is effective in reducing the gap between actual and minimum detection costs for a large set of possible applications, and is particularly useful for compensating global miscalibration effects such as those deriving from models that are not able to directly output verification log-likelihood ratios (e.g. Pairwise Support Vector Machines (PSVM) [9, 10], cosine scoring [11]), or those resulting from distribution mismatch between training and evaluation data. However, in many cases scores present also *trial-dependent* miscalibration, i.e. trials are differently affected by miscalibration sources, so that trial-dependent calibration transformations are required to obtain well-calibrated LLRs. In these cases trial-dependent calibration allows us to increase the discrimination capabilities of the calibrated scores, reducing both the actual *and* the minimum detection costs. A well-known example is utterance duration: global calibration models that are effective for long utterances usually are not effective for short segments and vice versa, whereas calibration transformations that depend on the duration of the trial utterances usually lead to improved minimum and actual costs [12, 13, 14, 15]. To address these issues, global calibration approaches have been extended to partially handle trial-dependent miscalibration. Prior-weighted Logistic Regression has been modified to incorporate side-information consisting, for example, in utterance duration and noise levels [12, 13, 14]. Utterance duration has also been effectively incorporated in generative approaches [15], and the latter work provides a framework that may be extended to handle also different sources of trial-level miscalibration. While effective, these methods require an explicit model for between-trial variability. Recently, an approach that extracts side-information directly from speaker embeddings has been proposed in [16]. The authors jointly train a discriminative Probabilistic Linear Discriminant Analysis [17] classifier and a calibration back-end that exploits a formulation that closely mimics the scoring expression of discriminative models [9, 10, 17] to produce a classifier that is robust to different evaluation conditions. While effective, the approach is, however, tied to a specific classification back-end.

As an alternative to trial-dependent calibration, score normalization [18, 19, 20, 21, 22, 23] can also be used to mitigate trial-level miscalibration [24, 25]. Score normalization aims at mapping the distribution of non-target scores of different enrollment speakers and test segments to fixed, common distributions (usually, a standard Gaussian). In contrast with calibration, however, score normalization tends to produce scores that are globally miscalibrated. A second step that performs global re-calibration is therefore usually required to produce good decisions. Furthermore, as we show in the next sections, while these methods are effective for some tasks, in some cases they may actually increase, rather than reduce, trial-level miscalibration, therefore lowering the discrimination power and the ver-

10.21437/Odyssey.2022-4

ification accuracy with respect to the original scores. In this work we address this issue by proposing a discriminative model based on prior-weighted logistic regression which employs, as side-information, the same kind of information that is used by score normalization approaches. The model is motivated by an analysis of the distribution of well-calibrated scores [3]. We show that our approach is able to address both global and trial-level miscalibration, with a negligible increase in complexity compared to discriminative global methods. We propose the model as an alternative to score normalization that provides well-calibrated results. It's worth noting that our approach is complementary to other trial-dependent models. For example, additional quality measures such as those used in [13] can be straightforwardly incorporated in our model whereas it's relatively easy to extend the work [16] to incorporate our score normalization-derived side-information.

The rest of the paper is organized as follows. In Section 2 we analyze the limitations of score normalization, showing that it may increase, rather than decrease, trial-level miscalibration. In Section 3 we show how we can incorporate impostor score statistics as quality measures for discriminative calibration based on logistic regression. Section 4 presents our experimental results. Conclusions are drawn in Section 5.

2. Score normalization and calibration

Score normalization aims at reducing trial-dependent miscalibration by normalizing the impostor (non-target) score distributions of different enrollment and/or test utterances. It employs (unlabeled) impostor cohorts to compute impostor scores for both the enrollment and test sides of a trial. Impostor score statistics are then used to normalize the trial score. Symmetric normalization (S-norm) [20] and Adaptive S-norm (AS-norm) [23, 22] are among the most successful approaches. Given a trial (e,t) and an impostor cohort[1] $\{x_i\}_{i=1}^N$ comprising N samples, S-norm computes the normalized score as the average between Z-normalized [18] and T-normalized [19] scores

$$s_{s\text{-}norm}(e,t) = \frac{1}{2}\left(\frac{s(e,t)-\mu(e)}{\sigma(e)} + \frac{s(e,t)-\mu(t)}{\sigma(t)}\right) \quad (1)$$

where $s(e,t)$ is the unnormalized score for the trial (e,t). The statistics

$$\mu(e),\ \sigma(e),\ \mu(t),\ \sigma(t) \quad (2)$$

are the mean and standard deviation of the set of impostor scores $\{s(e,x_i)\}_{i=1}^N$ and $\{s(x_i,t)\}_{i=1}^N$, respectively. Adaptive variants such as AS-norm [23, 22] introduce a cohort selection step. The statistics in (2) are computed from a subset of impostors, rather than from the full cohort. The subset is usually selected so that the cohort used to compute the enrollment statistics $\mu(e),\sigma(e)$ is similar to the test utterance t, and vice-versa. This allows for heterogeneous cohorts that can cover a wider range of possible conditions, and often (but not always) results in better accuracy with respect to non adaptive methods.

An advantage of score normalization is that it can leverage unlabeled datasets with just a single repetition per impostor speaker, and does not require additional side-information to be provided. On the contrary, current calibration approaches can exploit unlabeled datasets to estimate calibration parameters, but they still require multiple repetitions per speaker for global calibration, and external side-information in order to improve trial-level miscalibration. Score normalization, however, has two main drawbacks. First, it's not able to provide good global calibration, since it does not have access to information on the distribution of target scores. Global calibration is therefore still necessary to obtain well-calibrated, normalized scores. An additional dataset that includes multiple speaker repetitions is thus required also in this case to estimate the calibration parameters. The second drawback is that score normalization may, depending on the task, increase, rather than reduce, trial-level miscalibration. To show this, we consider a simple scenario where we fix an enrollment speaker i, and we assume that impostor and target scores are generated by Gaussian-distributed R.V.s, as in Figure 1. We also assume that the scores are well-calibrated LLRs. From [3], this requires that the same-speaker and different speaker distribution densities $f_{\mathfrak{S}}$ and $f_{\mathfrak{D}}$ are related by

$$f_{\mathfrak{D}}(s) = \mathcal{N}\left(s\left|-\frac{1}{2}v_i, v_i\right.\right), \quad f_{\mathfrak{S}}(s) = \mathcal{N}\left(s\left|\frac{1}{2}v_i, v_i\right.\right) \quad (3)$$

where v_i is a variance term. It is worth noting that the variance v_i affects the discriminability of the scores of the speaker: larger variance implies easier trials, while smaller variance implies harder trials. Since in practical use cases different enrollment utterances convey different amounts of information (e.g. because of different duration), it's reasonable to expect that different enrolled speakers will present score distributions that have different variances, even for perfectly calibrated verification models. We therefore consider a set where we have three different enrolled speakers, whose score distributions have different variance, as shown in Figure 2. Symmetric density pairs refer to the same enrollment speaker (and have the same variance). For each speaker, the corresponding scores are well calibrated by construction. If we assume a uniform probability that a trial will be formed for any of the three enrolled speakers, regardless of the class hypothesis, then the pooled scores for the

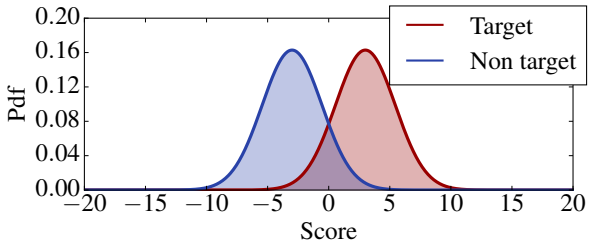


Figure 1: Gaussian-distributed score densities for a single speaker.

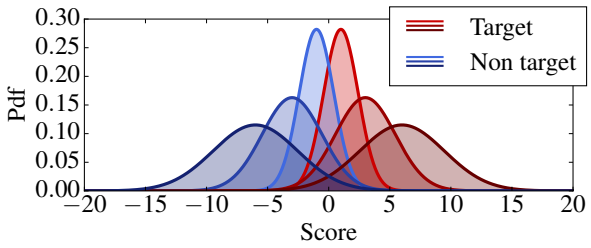


Figure 2: Gaussian-distributed score densities for three different speakers. Symmetric densities (densities with the same variance) correspond to the same speaker.

[1] In general, enrollment and impostor cohort may be different. To ease the exposition, we here consider a single cohort.

three speakers are also perfectly calibrated.

We now consider applying Z-normalization: the scores of each speaker are normalized so that the impostor densities become standard Gaussian. The results are shown in Figure 3. As we can see, the impostor score densities overlap, whereas the target densities have the same scale, but different locations. If we consider the pooled scores, it's easy verifying that the scores are no more globally well-calibrated, and equation (3) does not hold anymore. Furthermore, we can also observe that, in general, even if we are able to find a threshold that is optimal for a given speaker, the same threshold would be sub-optimal for the other two speakers, since the target densities have the same shape but different location. Indeed, score normalization has introduced speaker-dependent miscalibration. This is confirmed by the Bayes error plots in Figure 4, which plot minimum costs[2] for the original and Z-normalized scores. We can observe that the original scores result in lower costs over the whole range of application priors.

Summarizing, while score normalization may be effective in applications with significant trial-dependent miscalibration, for some tasks it may rather adversely affect performance. Furthermore, a global calibration step is nevertheless required to transform normalized scores in well-calibrated LLRs. To address the first issue, we propose an alternative approach that directly integrates the impostor score statistics used by score normalization with a global discriminative calibrator.

3. Impostor statistics as quality measures

In this section we show how impostor score statistics can be integrated into standard prior-weighted logistic regression as a particular kind of side-information, that can be linearly combined with the original scores to improve trial-dependent miscalibration while obtaining, at the same time, good global calibration. We start our derivations by considering again a Z-norm-style approach, in which we re-calibrate the scores of a set of given enrollment speakers with a speaker-dependent transformation. We will then extend the method to handle both sides of a trial. We assume that, given a speaker i, the scores are samples of Gaussian[3] distributed R.V.s $X_{\mathfrak{D},i}$, $X_{\mathfrak{S},i}$ that are well calibrated up to a speaker-dependent affine transformation:

$$X_{\mathfrak{D},i} \sim a_i X^{cal}_{\mathfrak{D},i} + b_i, \quad X_{\mathfrak{S},i} \sim a_i X^{cal}_{\mathfrak{S},i} + b_i, \tag{4}$$

where

$$X^{cal}_{\mathfrak{D},i} \sim \mathcal{N}\left(-\frac{1}{2}v_i, v_i\right), \quad X^{cal}_{\mathfrak{S},i} \sim \mathcal{N}\left(\frac{1}{2}v_i, v_i\right). \tag{5}$$

The term v_i represents the variance of the well-calibrated R.V.s $X^{cal}_{\mathfrak{S},i}$ and $X^{cal}_{\mathfrak{D},i}$, while a_i and b_i are enrollment-dependent miscalibration parameters. If we knew both a_i and b_i for each enrolled speaker, we would be able to re-calibrate the speaker scores through the transformation

$$s_{cal} = \frac{s - b_i}{a_i}. \tag{6}$$

We can observe that, if we assume $a_i = 1$ for all speakers, then b_i can be computed once we know the mean $m_{\mathfrak{D},i}$ and variance

[2]We compute minimum costs since Z-normalized scores would otherwise require a further score calibration step.

[3]Although we have shown that the Gaussian assumption may be inaccurate for some tasks [6, 15], in this context it's sufficient for our goals and allows us to greatly simplify the derivations.

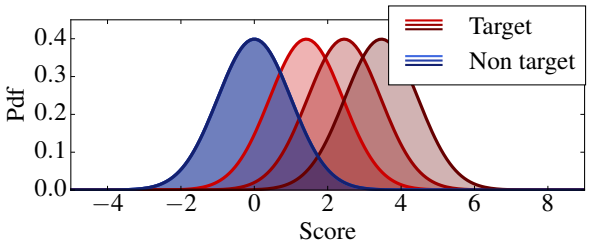


Figure 3: Score densities for the three speakers after Z-norm. The non-target score densities (in blue) overlap.

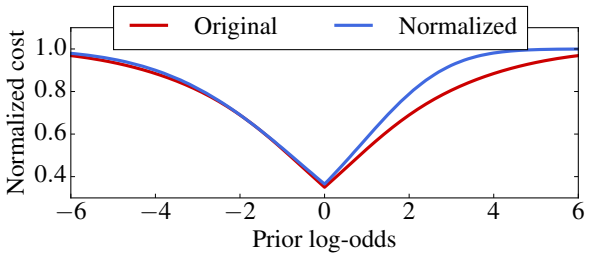


Figure 4: Bayes error plots for the synthetic scores of Figures 2 (in red) and 3 (in blue).

v_i of $X_{\mathfrak{D},i}$:

$$b_i = m_{\mathfrak{D},i} + \frac{1}{2}av_i = m_{\mathfrak{D},i} + \frac{1}{2}v_i. \tag{7}$$

Although in practice also $m_{\mathfrak{D},i}$ and v_i are unknown, we may replace them with an estimate obtained from the set of impostor scores for speaker i. In the general case, however, a_i is unknown, and we are not able to estimate both a_i and b_i without knowledge of the target score distribution for speaker i. Score normalization addresses this issue by further assuming that the impostor scores of each speaker have been generated by R.V.s which have the same distribution, up to a speaker-dependent affine transformation.

$$X_{\mathfrak{D},i} \sim a_i X^{cal}_{\mathfrak{D}} + b_i, \quad X^{cal}_{\mathfrak{D}} \sim \mathcal{N}\left(-\frac{1}{2}v, v\right). \tag{8}$$

For score normalization to be effective, the normalized scores should be well-calibrated for each speaker, which requires that

$$X_{\mathfrak{S},i} \sim a_i X^{cal}_{\mathfrak{S}} + b_i, \quad X^{cal}_{\mathfrak{S}} \sim \mathcal{N}\left(\frac{1}{2}v, v\right). \tag{9}$$

The calibrated distributions are thus assumed to have the same mean and the same variance $v_i = v$ for all speakers. This allows estimating both terms a_i and b_i up to speaker-independent (global) factors, which can then be estimated by a global linear calibration method. Score normalization can thus be close to optimal for datasets that closely match the same-variance assumption. However, as we have shown in the previous section, this assumption may not be very accurate in real use cases, and may result in an increase of trial-dependent miscalibration compared to the original scores.

In this work we replace the same-variance assumption $v_i = v$ with the dual assumption that the scaling factors a_i are independent of the enrollment speaker, i.e. $a_i = a$ for any speaker.

This corresponds to the assumption that most trial-dependent miscalibration can actually be modeled as a trial-dependent (or speaker dependent, in a Z-norm-style scenario) shift, while the variance variability is actually due to the intrinsic differences in the amount of information provided by different utterances. We therefore assume that the miscalibrated scores are generated by R.V.s with distributions

$$X_{\mathfrak{D},i} \sim aX^{cal}_{\mathfrak{D},i} + b_i, \quad X_{\mathfrak{S},i} \sim aX^{cal}_{\mathfrak{S},i} + b_i, \tag{10}$$

where $X^{cal}_{\mathfrak{D},i}$ and $X^{cal}_{\mathfrak{S},i}$ are the calibrated R.V.s as defined in (5). Putting together (10) and (5) we obtain the model

$$X_{\mathfrak{D},i} \sim \mathcal{N}\left(m_{\mathfrak{D},i}, v_{\mathfrak{D},i}\right), \quad X_{\mathfrak{S},i} \sim \mathcal{N}\left(m_{\mathfrak{S},i}, v_{\mathfrak{S},i}\right), \tag{11}$$

where $m_{\mathfrak{S},i}, m_{\mathfrak{D},i}, v_{\mathfrak{S},i}$ and $v_{\mathfrak{D},i}$ are the parameters of the target and non-target scores, tied as

$$m_{\mathfrak{D},i} = -\frac{1}{2}av_i + b_i;\ m_{\mathfrak{S},i} = \frac{1}{2}av_i + b_i;\ v_{\mathfrak{D},i} = v_{\mathfrak{S},i} = a^2 v_i. \tag{12}$$

From (10), the transformation that maps an uncalibrated score s of speaker i to a calibrated score s_{cal} is simply given by

$$s_{cal} = \frac{s - b_i}{a}. \tag{13}$$

From (12) it follows that $b_i = m_{\mathfrak{D},i} + \frac{1}{2}\frac{v_{\mathfrak{D},i}}{a}$, thus we can express s_{cal} in terms of $a, m_{\mathfrak{D},i}$ and $v_{\mathfrak{D},i}$ as

$$s_{cal} = \frac{1}{a}s - \frac{1}{a}m_{\mathfrak{D},i} - \frac{1}{2a^2}v_{\mathfrak{D},i}. \tag{14}$$

As for score normalization, we can replace the unknown terms $m_{\mathfrak{D},i}$ and $v_{\mathfrak{D},i}$ with the estimates obtained from an impostor cohort. However, we still need to estimate the scaling term a. This requires knowledge of the target score distribution, and thus (labeled) calibration data. Although we may proceed with a Maximum Likelihood approach, we observe that the calibration transformation (14) corresponds to a linear calibration model with a speaker-dependent shift. In particular, we can interpret the terms $m_{\mathfrak{D},i}$ and $v_{\mathfrak{D},i}$ as *side-information* (quality measures) [13] that is linearly combined with a re-scaled score. We can therefore employ discriminative linear fusion models such as prior-weighted logistic regression [1] to estimate the calibration transformation. Furthermore, we can extend the model so that the relevance of the different terms is automatically estimated from the data. We assume a calibration model, augmented with impostor side-information, given by

$$s_{cal} = \alpha s + \beta m_{\mathfrak{D},i} + \gamma v_{\mathfrak{D},i} + k, \tag{15}$$

where α, β, γ, k are estimated by training a prior-weighted logistic regression model over a set of labeled data. The model presents two advantages: (i) normalization and calibration are jointly estimated to optimize a proper scoring rule, and (ii) in contrast with standard score normalization, since we estimate weights β and γ, our model can avoid introducing trial-dependent miscalibration (e.g. estimating $\beta \approx \gamma \approx 0$) for scenarios that do not closely match our modeling assumptions.

We can trivially extend the model (14) to symmetrically handle both enrollment and test statistics, by further incorporating the test statistics as additional side-information. We consider again a trial (e, t), and we let m_e, v_e, m_t, v_t be the mean and variance of scores obtained comparing the enrollment and test utterances, respectively, against the normalization cohort(s). The calibration function becomes

$$s_{cal}(e,t) = \alpha s(e,t) + \beta m_e + \gamma v_e + \delta m_t + \epsilon v_t + k. \tag{16}$$

In practice, we have observed experimentally that we can further improve the accuracy of the model by considering also interactions between the test and the enrollment impostor distributions. The final model we propose includes an additional term:

$$s_{cal}(e,t) = \alpha s(e,t) + \beta m_e + \gamma v_e + \delta m_t + \epsilon v_t + \zeta\sqrt{v_e v_t} + k. \tag{17}$$

All model parameters $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta, k)$ can be estimated by means of prior-weighted logistic regression. We observe that additional side-information, such as utterance duration [13], can also be trivially incorporated in our model. At the same time, it's also straightforward to integrate our impostor statistics as side information in alternative calibration approaches, such as [16]. Furthermore, adaptive score normalization approaches have proven to be more effective in scenarios where the impostor cohort is heterogeneous. We can easily extend our approach to incorporate adaptive cohort selection by replacing the statistics in (17) with those computed from the selected cohort subsets for each trial.

4. Experimental results

In this section we analyze the performance of our approach on the SITW [26] and SRE 2019 [27] datasets. We contrast our method with global calibration based on prior-weighted logistic regression, and with S-norm and its adaptive variant AS-norm, implemented[4] as in [24]. We refer to our approach as C-norm. We also consider an adaptive variant, Adaptive C-norm (AC-norm). The cohort selection strategy is the same used for AS-norm (the cohort subsets are the same for the two approaches).

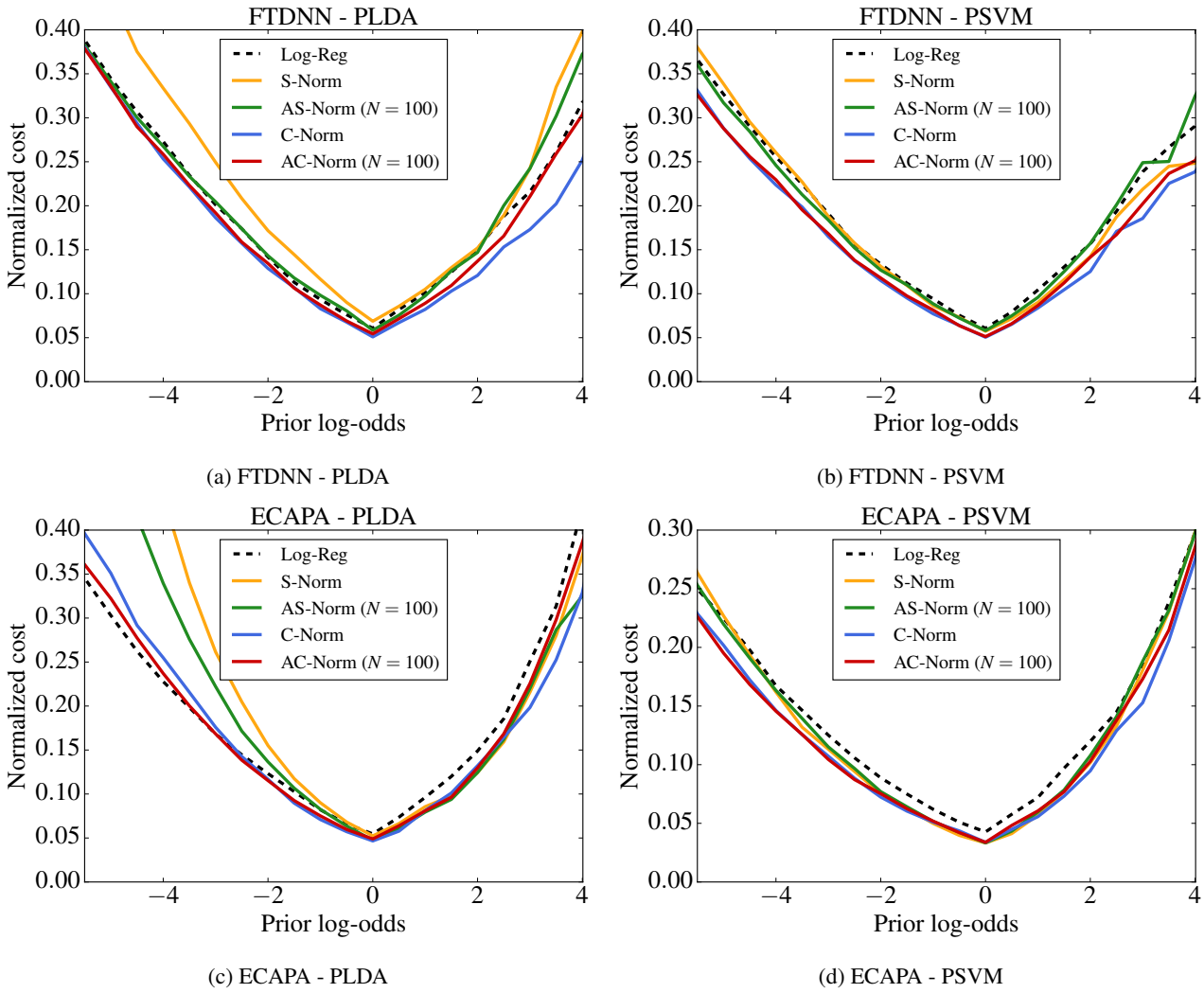
4.1. Embedding extractors and classification backends

We tested two different speaker embedding extractors, trained with data from Mixer6, NIST 2004–2010, Switchboard, VoxCeleb1, and the development set of VoxCeleb2. The first extractor is based on a Factorized Time-Delay Neural Network (FTDNN) [28], implemented as in [29]. The network has been trained for 20 epochs using softmax and cross-entropy loss on clean data. Embeddings are 512-dimensional. The second network is based on the ECAPA architecture [30]. The network has been trained for 10 epochs using Additive Angular Margin softmax [31, 32] and cross-entropy loss. The MUSAN [33] and the AIR [34] datasets were used to augment the training data with music, noise, babble and reverberation. Embedding dimensionality was set to 192. As backends, we consider both Probabilistic Linear Discriminant Analysis (PLDA) [35, 20] and Pairwise Support Vector Machine (PSVM) [9, 10, 36, 37]. Both classifiers have been trained with data from Mixer6, NIST 2004–2010, Switchboard, VoxCeleb1 and NIST SRE 2018 evaluation data. FTDNN embeddings were reduced to 200 (PLDA) or 400 (PSVM) dimensions by means of LDA. Whitening and length-normalization were applied for both front-ends. For PSVM embeddings were further processed by Within Class Covariance Normalization (WCCN).

4.2. Evaluation metrics

Results are reported in terms of Cost of Log-Likelihood Ratio C_{llr} [38, 39, 40] and of primary metrics C^{SITW}_{prim} and C^{SRE19}_{prim} defined for each task. Additionally we report Equal Error Rate

[4] We adopt the AS-norm2 variant of [24]. AS-norm2 is slightly different from our original proposal [23], however the two models provide almost identical results, and the former requires lower computational resources.



(a) FTDNN - PLDA

(b) FTDNN - PSVM

(c) ECAPA - PLDA

(d) ECAPA - PSVM

Figure 5: Results for different embedding extractors and different classifiers on the SITW Evaluation dataset. The AS-norm cohort size ($N = 100$) was selected according to the results on the calibration set. For C-norm the best results are obtained with the full cohort, both on the calibration and evaluation sets. As reference, the results with adaptive selection ($N = 100$) are given also for AC-norm.

(EER) and actual detection costs as defined for NIST SRE 2008 (DCF08 in the tables). We also provide normalized Bayes error rate [39] plots that show the normalized Detection Cost Function (DCF) corresponding to different target prior log-odds.

4.3. Data organization

4.3.1. SITW

For SITW experiments the calibration models have been trained on the SITW Development set. Results are reported on the SITW Evaluation set. Since SITW contains some speakers that are also in the VoxCeleb datasets that were used to train the embedding extractors, we removed these speakers from the development and evaluation sets. The development set was split into two, non overlapping, parts. The first part, comprising about 30% of the speakers, was used as normalization cohort. The second part was used to estimate the calibration parameters for all models.

4.3.2. SRE 2019

For SRE 2019 experiments calibration models have been trained on the SRE 2019 Progress set. The impostor cohort consists of the unlabeled portion of SRE 2018 development data [41]. Results are reported on the SRE 2019 Evaluation dataset.

4.4. Results

4.4.1. SITW

The first set of experiments compares S-norm and its adaptive variant AS-norm with C-norm on the SITW dataset. For AS-norm, the cohort size $N = 100$ was selected as the optimal size based on the results obtained on the development set. For C-norm best results were obtained in all cases using the full cohort set. Nevertheless, we also compare the performance of AC-norm using the same cohort size $N = 100$. S-norm and AS-norm scores have been re-calibrated by means of prior-weighted logistic regression estimated on the development set. All the calibration models have been trained with a target prior

Table 1: Results for different embedding extractors and different classifiers on the SITW Evaluation dataset. Minimum costs are reported for the *unnormalized* scores. LogReg rows correspond to global calibration.

(a) FTDNN - PLDA

	EER	DCF08	C_{prim}^{SITW}	C_{llr}
Min cost	*3.0%*	*0.159*	*0.312*	*0.110*
Log-Reg	3.0%	0.161	0.316	0.113
S-norm	3.5%	0.194	0.389	0.130
AS-norm ($N = 100$)	3.0%	0.161	0.305	0.116
C-norm	**2.5%**	**0.145**	**0.298**	**0.098**
AC-norm ($N = 100$)	2.7%	0.149	0.299	0.104

(b) FTDNN - PSVM

	EER	DCF08	C_{prim}^{SITW}	C_{llr}
Min cost	*3.0%*	*0.144*	*0.292*	*0.109*
Log-Reg	3.0%	0.145	0.295	0.113
S-norm	2.8%	0.146	0.301	0.107
AS-norm ($N = 100$)	2.8%	0.141	0.290	0.110
C-norm	**2.5%**	0.130	**0.259**	**0.095**
AC-norm ($N = 100$)	**2.5%**	**0.126**	0.262	0.098

(c) ECAPA - PLDA

	EER	DCF08	C_{prim}^{SITW}	C_{llr}
Min cost	*2.7%*	*0.132*	*0.269*	*0.105*
Log-Reg	2.7%	0.135	0.272	0.111
S-norm	2.6%	0.182	0.574	0.115
AS-norm ($N = 100$)	2.4%	0.156	0.433	0.103
C-norm	**2.3%**	0.131	0.303	**0.096**
AC-norm ($N = 100$)	2.5%	**0.128**	**0.285**	0.102

(d) ECAPA - PSVM

	EER	DCF08	C_{prim}^{SITW}	C_{llr}
Min cost	*2.1%*	*0.097*	*0.198*	*0.079*
Log-Reg	2.1%	0.098	0.202	0.083
S-norm	**1.7%**	0.087	0.200	0.072
AS-norm ($N = 100$)	**1.7%**	0.090	0.197	0.073
C-norm	**1.7%**	**0.081**	0.176	**0.069**
AC-norm ($N = 100$)	**1.7%**	0.084	**0.173**	0.072

equal to 0.1. Figure 5 shows the Bayes error plots for the different front-end/back-end combinations. Table 1 reports minimum costs computed on the original, unnormalized scores, as well as the EER and actual costs of the different score normalization and calibration methods.

We can observe that, for PLDA based systems (Figures 5a and 5c, Tables 1a and 1c), S-norm provides significantly worse results with respect to the other approaches, and in some cases significantly degrades performance compared to unnormalized scores. On the contrary, C-norm with the full cohort set is able to significantly improve the results with respect to unnormalized scores (LogReg row). Surprisingly, even though the full cohort set is small, cohort selection is effective for AS-norm. A cohort set of size $N = 100$ provides better results, even though it still incurs in a degradation in terms of C_{llr} with respect to unnormalized scores. On the other hand, as we were expecting given the cohort set size, adaptive cohort selection is not effective for C-norm in this scenario. AC-norm models have slightly worse C_{llr} than C-norm models, although we can observe an improvement in terms of primary cost for the ECAPA-based system.

PSVM models (Figures 5b and 5d, Tables 1b and 1d) provide, in general, better results than PLDA models, especially with the ECAPA frontend. S-norm is much more effective in this case, with results that are similar to those of AS-norm with $N = 100$. As for PLDA, however, C-norm with the full cohort set provides the best results in terms of C_{llr}. For FTDNN embeddings we can observe a significant improvement compared to both S-norm and raw scores. For the ECAPA-based models the relative improvement is smaller, although we can observe a significant improvement in terms of actual primary cost.

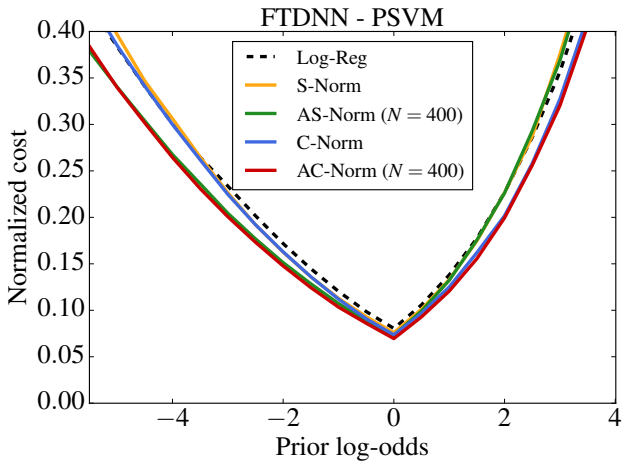
Overall, C-norm proves to be effective, consistently outperforming both global calibration, S-norm and AS-norm.

4.4.2. SRE 2019

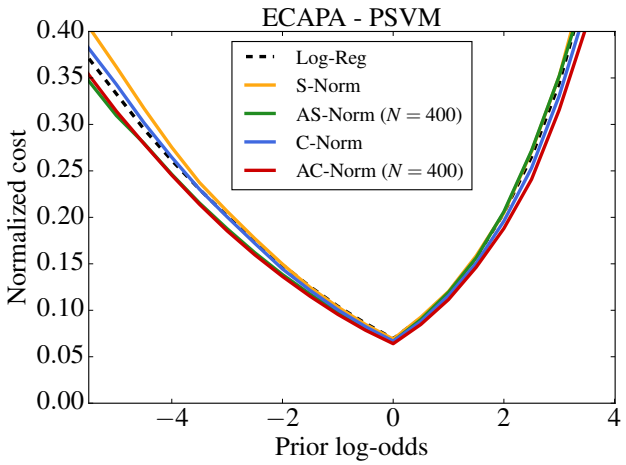
In the second set of experiments we compare the results of S-norm and AS-norm with C-norm and AC-norm on SRE 2019. We concentrate on PSVM models, since the unnormalized results are significantly better than those of PLDA-based systems. The results are shown in Figure 6 and Table 2. In contrast with SITW results, for this task we can observe that adaptive models perform consistently better than their non-adaptive counterparts. This was expected, since the cohort size is much larger and probably more heterogeneous. Comparing the non adaptive models, we can see that C-norm performs better than S-norm. Although the improvement in terms of C_{llr} is small, we can observe that C-norm is consistently better than S-norm in the low false reject region, whereas it achieves similar or slightly better performance in the low false alarm region. Similar considerations hold for the adaptive models. AC-norm provides small but consistent improvements in terms of C_{llr} with respect to AS-norm, is consistently better in the low false reject region and achieves similar performance in the low false alarm region.

5. Conclusions

We have presented a novel approach for trial-dependent score calibration that incorporates impostor score statistics that are usually employed by score normalization methods as additional side information that can be linearly combined with the original score. The proposed approach overcomes some of the limitations of score normalization, improving the performance of the verification systems over a wide range of operating points. Our method can be trivially implemented using existing toolkits (e.g. [42]) and can be trivially extended to incorporate additional side information or integrated with recent, condition-robust discriminative approaches. In the future we will analyze



(a) FTDNN - PSVM



(b) ECAPA - PSVM

Figure 6: Results for different embedding extractors and PSVM classifier on the SRE 2019 Evaluation dataset. The AS-norm and AC-norm cohort size ($N = 400$) was selected according to the results on the calibration set.

Table 2: Results for different embedding extractors and PSVM classifier on the SRE 2019 Evaluation dataset. Minimum costs are reported for the *unnormalized* scores. LogReg rows correspond to global calibration.

(a) FTDNN - PSVM

	EER	DCF08	C_{prim}^{SRE19}	C_{llr}
Min cost	*4.0%*	*0.188*	*0.374*	*0.153*
Log-Reg	4.0%	0.188	0.380	0.158
S-norm	3.8%	0.178	0.392	0.155
AS-norm (400)	3.7%	0.166	0.337	0.148
C-norm	3.7%	0.180	0.381	0.149
AC-norm (400)	**3.5%**	**0.162**	**0.336**	**0.142**

(b) ECAPA - PSVM

	EER	DCF08	C_{prim}^{SRE19}	C_{llr}
Min cost	*3.5%*	*0.162*	*0.326*	*0.136*
Log-Reg	3.5%	0.164	0.329	0.145
S-norm	3.5%	0.166	0.356	0.147
AS-norm (400)	3.4%	0.152	**0.307**	0.137
C-norm	3.3%	0.161	0.337	0.142
AC-norm (400)	**3.2%**	**0.150**	0.310	**0.134**

possible methods to extend these results to generative calibration models.

6. References

[1] N. Brummer 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.

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

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

[4] 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.

[5] 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.

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

[7] 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)*, 2019, pp. 6121–6125.

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

[9] 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.

[10] 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.

[11] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, and P. Ouellet, "Support Vector Machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Interspeech 2009*, 2009, pp. 1559–1562.

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

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

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

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

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

[17] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, and N. Brümmer, "Discriminatively trained Probabilistic Linear Discriminant Analysis for speaker verification," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4832–4835.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 31–44, 2000.

[19] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 4254, 2000.

[20] P. Kenny, "Bayesian speaker verification with Heavy-Tailed Priors," in *Odyssey 2010*, 2010.

[21] D. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker recognition," in *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[22] Z. Karam, W. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4512–4515.

[23] S. Cumani et al., "Comparison of speaker recognition approaches for real applications," in *Proceedings of Interspeech 2011*, 2011, pp. 2365–2368.

[24] P. Matějka et al., "Analysis of score normalization in multilingual speaker recognition," in *Interspeech 2017*, 2017.

[25] D. Colibro, C. Vair, E. Dalmasso, K. Farrell, G. Karvitsky, S. Cumani, and P. Laface, "Nuance - Politecnico di Torino's 2016 NIST Speaker Recognition Evaluation system," in *Interspeech 2017*, 2017.

[26] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The Speakers in the Wild (SITW) Speaker Recognition Database," in *Interspeech 2016*, 2016, pp. 818–822.

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

[28] D. Povey et al., "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech 2018*, 2018.

[29] J. Villalba et al., "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Comput. Speech Lang.*, vol. 60, 2020.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, 2020.

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

[32] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," *APSIPA ASC 2019*, pp. 1652–1656, 2019.

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

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

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

[36] S. Cumani, N. Brümmer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i–vector space," in *Proceedings of ICASSP 2011*, 2011, pp. 4852–4855.

[37] S. Cumani et al., "Gender independent discriminative speaker recognition in i-vector space," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[38] 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.

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

[40] 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, 2007.

[41] "The NIST 2018 speaker recognition evaluation plan," 2018, Available at https://www.nist.gov/system/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf.

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