POLITECNICO DI TORINO
Repository ISTITUZIONALE

From adaptive score normalization to adaptive data normalization for speaker verification systems

Original
From adaptive score normalization to adaptive data normalization for speaker verification systems / Cumani, S., Sarni, S.. - ELETTRONICO. - (2023), pp. 5296-5300. (INTERSPEECH 2023 Dublin (IE) 20th - 24th August 2023) [10.21437/Interspeech.2023-266].

Availability:
This version is available at: 11583/2979929 since: 2023-07-05T11:55:54Z

Publisher:
ISCA

Published
DOI:10.21437/Interspeech.2023-266

Terms of use:

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

Publisher copyright

(Article begins on next page)

26 November 2026

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

From adaptive score normalization to adaptive data normalization for speaker verification systems

Sandro Cumani, Salvatore Sarni

Politecnico di Torino, Italy

sandro.cumani@polito.it, salvatore.sarni@polito.it

Abstract

Domain and trial-dependent mismatch between training and evaluation data can severely affect the performance of speaker verification systems, and are usually addressed either at embedding level, with methods that try matching the distribution of in-domain and out-of-domain data, or at score level by means of calibration and score normalization approaches. In this work we propose an alternative to score normalization that leverages the adaptive cohort selection of Adaptive S-norm (AS-norm), but performs normalization at embedding rather than at score level. Experimental results on SRE 2016 and SRE 2019 show that the proposed method is able to outperform other approaches in presence of severe mismatch, and achieves similar performance in scenarios where score normalization is less important. Furthermore, in contrast with AS-norm, our approach allows independently normalizing the enrollment and test segments, and has negligible computational cost at scoring time.

Index Terms: speaker recognition, score normalization, adaptive score normalization, speaker embeddings.

1. Introduction

Off-the-shelf speaker verification systems are typically required to cope with scenarios characterized by significant mismatch between the use-case data and the data employed to train the recognizer. The domain mismatch is often coupled with trial-level mismatch, i.e. mismatch due to different utterances being characterized by different nuisance factors such as duration, noise type, noise level or similar. These two factors can significantly decrease the performance of the recognizer. Several approaches have been proposed to address domain mismatch, either by adapting the classification models [1, 2] or by applying a transformation to the evaluation data [3, 4, 5, 6, 7] to reduce the distribution mismatch between evaluation and training populations. These transformations usually try to match the population means [7] and, in some cases, also the corresponding covariance matrices [4, 5, 6]. The transformed data can then be classified with the off-the-shelf recognizer, trained on out-of-domain data. While effective, these methods do not address trial-level mismatch, i.e. they apply a single, trial-independent transformation to the data. Alternative to domain compensation approaches, score normalization [8, 9, 10, 11, 12, 13] has proven to be effective at compensating both domain and trial-level mismatch [14, 15, 16, 17]. Score normalization techniques employ an unlabeled normalization set to estimate impostor score statistics that are used to normalize the impostor score distribution for a given trial. Particularly successful are adaptive methods such as Adaptive S-norm (AS-norm) [11, 12], which combine score normalization with a cohort selection strategy that aims at identifying impostor utterances that are similar to the trial at hand. This allows obtaining well-matched cohort sets and provides significant improvements with respect to methods that employ a fixed impostor cohort.

In this work we investigate an approach that exploits the effective cohort selection strategy of AS-norm to perform normalization at embedding, rather than score level. In particular, we propose to employ an adaptive cohort, estimated as in AS-norm, to estimate an utterance-dependent mean vector in embedding space. The computed mean vector is then used to normalize a speaker embedding, and the out-of-domain classifier is employed to classify normalized trials. Our approach is similar to the work [7], which also proposes an utterance-dependent mean compensation. In [7] the authors employ an ad-hoc back-end to assess whether utterances belong to similar conditions, a condition being a set of characteristics such as noise type, noise level, gender and similar. The condition back-end is used to select an impostor set to compute utterance-dependent mean vectors, which, as in our case, are used to normalize the evaluation population. Our work differs from [7] in that we employ a selection strategy that is derived from score normalization techniques. Our approach therefore does not require training a condition-similarity back-end, which may be difficult in case conditions are not well defined even for the training population, or when conditions significantly differ between training and evaluation data. On the contrary, our selection mechanism directly exploits score similarities to automatically extract impostor sets that properly characterize an utterance, and is able to significantly improve recognition performance, as confirmed by our experimental results.

The work is organized as follows. Section 2 recalls AS-norm and its cohort selection mechanism. Section 3 presents our model, highlighting the relationship with AS-norm and providing an intuitive explanation behind the use of AS-norm cohort selection to compute the utterance-dependent means required by our approach. Section 4 illustrates our experimental results. Finally, Section 5 presents our conclusions.

2. Adaptive score normalization

Let $(\mathbf{e}, \mathbf{t})$ represent a trial composed of an enrollment segment $\mathbf{e}$ and a test segment $\mathbf{t}$. Typically, $\mathbf{e}$ and $\mathbf{t}$ are the speaker embeddings for the enrollment and test audio. Standard back-ends such as Probabilistic Linear Discriminant Analysis (PLDA) [18, 14] and Pairwise Support Vector Machines (PSVM) [19, 20] allow computing a score $s(\mathbf{e}, \mathbf{t})$ that, after a calibration step, can be interpreted as a log-likelihood ratio between the target and non-target trial hypotheses. In presence of severe mismatch between the training and evaluation population and differences in the characteristics of the evaluation utterances, however, global calibration is usually not sufficient

10.21437/Interspeech.2023-266

to obtain good performance. In these cases, score normalization can improve performance by partially compensating both dataset and intra-trial mismatch [15, 16, 17]. The most common techniques for score normalization are derived from Symmetric normalization (S-norm) [21], that computes a normalized score as

$$s_{snorm}(\mathbf{e},\mathbf{t}) = \frac{s(\mathbf{e},\mathbf{t}) - \mu(\mathbf{e})}{2\sigma(\mathbf{e})} + \frac{s(\mathbf{e},\mathbf{t}) - \mu(\mathbf{t})}{2\sigma(\mathbf{t})} \ , \qquad (1)$$

where $\mu(\mathbf{e})$, $\sigma(\mathbf{e})$, $\mu(\mathbf{t})$, $\sigma(\mathbf{t})$ are the mean and standard deviation of the set of impostor scores $\{s(\mathbf{e},\mathbf{x}_i)\}_{i=1}^N$ and $\{s(\mathbf{x}_i,\mathbf{t})\}_{i=1}^N$, respectively, computed from a set of N (unlabeled) impostor segments $\{\mathbf{x}_1 \dots \mathbf{x}_N\}$. In [11] and [12] S-norm was extended to incorporate an adaptive cohort selection step derived from Adaptive T-norm [10]. Adaptive S-norm computes utterance dependent cohort sets $C(\mathbf{e})$ and $C(\mathbf{t})$ that contain utterances that are similar to the enrollment or test segments, respectively, and employs these cohort sets to compute the statistics required to normalize the trial score. Let

$$\begin{aligned} \mathbf{v}_t &= \begin{bmatrix} s(\mathbf{x}_1,\mathbf{t}) & s(\mathbf{x}_2,\mathbf{t}) & \dots & s(\mathbf{x}_N,\mathbf{t}) \end{bmatrix} \\ \mathbf{v}_i &= \begin{bmatrix} s(\mathbf{x}_1,\mathbf{x}_i) & s(\mathbf{x}_2,\mathbf{x}_i) & \dots & s(\mathbf{x}_N,\mathbf{x}_i) \end{bmatrix} \end{aligned} \qquad (2)$$

be the vectors of scores of test segment $\mathbf{t}$ and of normalization segments $\mathbf{x}_i$ w.r.t the whole impostor set. The cohort set $C(\mathbf{t})$ is selected by computing the squared Euclidean distance[1]

$$d_i = \|\mathbf{v}_i - \mathbf{v}_t\|^2 \qquad (3)$$

between score vectors $\mathbf{v}_i$ and $\mathbf{v}_t$, and keeping the K impostor utterances corresponding to the lowest difference (typically, K ranges from 200 to 400). The set $C(\mathbf{t})$ contains utterances that are similar to the test segment, and is employed to compute the enrollment normalization statistics $\mu(\mathbf{e}|\mathbf{t})$ and $\sigma(\mathbf{e}|\mathbf{t})$, corresponding to the mean and standard deviations of the scores $\{s(\mathbf{e},\mathbf{x}_i)|\mathbf{x}_i \in C(\mathbf{t})\}$ obtained by comparing the enrollment segment with the utterances in $C(\mathbf{t})$. The test statistics $\mu(\mathbf{t}|\mathbf{e})$ and $\sigma(\mathbf{t}|\mathbf{e})$ are computed in a similar way, reversing the role of enrollment and test. Finally, the normalized score is given by

$$s_{asnorm}(\mathbf{e},\mathbf{t}) = \frac{s(\mathbf{e},\mathbf{t}) - \mu(\mathbf{e}|\mathbf{t})}{2\sigma(\mathbf{e}|\mathbf{t})} + \frac{s(\mathbf{e},\mathbf{t}) - \mu(\mathbf{t}|\mathbf{e})}{2\sigma(\mathbf{t}|\mathbf{e})} \qquad (4)$$

3. Adaptive data normalization

Although score normalization techniques are usually effective at reducing significant trial-level mismatch, it can be shown that these approaches are not optimal and can actually reduce performance for well-behaved (i.e. well-calibrated) scores. We refer to [17] for further analysis of this phenomenon. On the other hand, directly reducing the mismatch between training and evaluation *embeddings* distribution on a *per-trial* basis would naturally reduce dataset-level and trial-level mismatch by providing a more accurate match between the back-end model assumptions and the evaluation data. Similarly to Adaptive Mean normalization (AM-norm) [7], we propose to replace normalization at score level with a method that directly normalizes the enrollment and test embeddings by re-centering the embeddings with respect to a mean vector computed from a set of impostor utterances. In contrast with [7] we select an utterance-dependent impostor cohort using the same mechanism of AS-norm. In particular, for each enrollment and test segment we independently compute the corresponding cohort sets $C(\mathbf{e})$ and $C(\mathbf{t})$ as detailed in the previous section, and we independently normalize each embedding by removing the mean of the corresponding cohort

$$\begin{aligned} \widehat{\mathbf{e}} &= \mathbf{e} - \mathbf{m}_e \ , \quad \mathbf{m}_e = \frac{1}{|C(\mathbf{e})|} \sum_{i|\mathbf{x}_i \in C(\mathbf{e})} \mathbf{x}_i \ , \\ \widehat{\mathbf{t}} &= \mathbf{t} - \mathbf{m}_t \ , \quad \mathbf{m}_t = \frac{1}{|C(\mathbf{t})|} \sum_{i|\mathbf{x}_i \in C(\mathbf{t})} \mathbf{x}_i \ . \end{aligned} \qquad (5)$$

The normalized enrollment and test segments are then scored using the standard back-end, as

$$s_{adnorm} = s\left(\widehat{\mathbf{e}},\widehat{\mathbf{t}}\right) = s(\mathbf{e}-\mathbf{m}_e, \mathbf{t}-\mathbf{m}_t) \ , \qquad (6)$$

where $s(\cdot,\cdot)$ is the scoring function of the off-the-shelf recognizer. We refer to our approach as Adaptive Data normalization (AD-norm). In practice, many back-ends employ length normalization to pre-process the embeddings. In this case, we re-center the L2-normalized embeddings, but we also apply a second L2-normalization step to better match the classifier assumptions.

Our approach is able to automatically select the impostor segments that are similar to the trial segments, without requiring an ad-hoc condition classifier as in [7]. As shown in the next section, this allows for a significant improvement with respect to AM-norm. Since the normalization is applied independently to the enrollment and test embeddings, it can be computed at embedding extraction time, with no impact on scoring costs. Finally, we note that, in contrast with AS-norm, AD-norm can be effectively paired with the clustering feature framework of [22] for large-scale clustering of speaker vectors.

In the following we analyze the similarities between the AD-norm and AS-norm scoring functions to provide an intuitive justification for using the AS-norm selection criterion. In particular, we show that we can interpret part of the effects of AS-norm as a sub-optimal proxy of embedding-level normalization. We start consider a zero-mean PLDA (or PLDA-derived model) with scoring function [19, 20]

$$s(\mathbf{e},\mathbf{t}) = \mathbf{e}^T\mathbf{A}\mathbf{e} + \mathbf{t}^T\mathbf{A}\mathbf{t} + \mathbf{e}^T\mathbf{B}\mathbf{t} \qquad (7)$$

and we assume that the distribution of training and evaluation data is the same up to a shift of the evaluation mean. The optimal scoring strategy would then consist in re-centering the evaluation data. Let $\mathbf{m}_e$ and $\mathbf{m}_t$ be the mean vectors of the distributions of the enrollment and test segments, respectively. The optimal score for trial $(\mathbf{e},\mathbf{t})$ would be

$$\begin{aligned} s_{opt}(\mathbf{e},\mathbf{t}) &= s(\mathbf{e}-\mathbf{m}_e, \mathbf{t}-\mathbf{m}_t) \\ &= (\mathbf{e}-\mathbf{m}_e)^T\mathbf{A}(\mathbf{e}-\mathbf{m}_e) + (\mathbf{t}-\mathbf{m}_t)^T\mathbf{A}(\mathbf{t}-\mathbf{m}_t) \\ &\quad + (\mathbf{e}-\mathbf{m}_e)^T\mathbf{B}(\mathbf{t}-\mathbf{m}_t) \ , \end{aligned} \qquad (8)$$

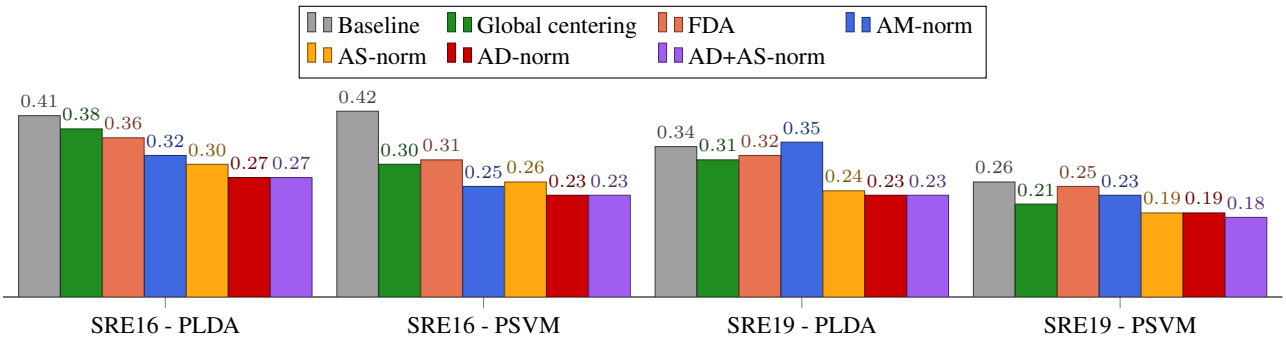
For the AD-norm approach we would replace $\mathbf{m}_e$ and $\mathbf{m}_t$ with the estimates obtained from the impostor sets $C(\mathbf{e})$ and $C(\mathbf{t})$. We now consider the scoring function of AS-norm, restricted to centering of the scores (i.e., we ignore the variance normalization of AS-norm):

$$s_{asnorm}(\mathbf{e},\mathbf{t}) = \frac{1}{2}(s_t(\mathbf{e},\mathbf{t}) - \boldsymbol{\mu}(\mathbf{e}|\mathbf{t})) + \frac{1}{2}(s_t(\mathbf{e},\mathbf{t}) - \boldsymbol{\mu}(\mathbf{t}|\mathbf{e})) \qquad (9)$$

where the score statistics $\boldsymbol{\mu}(\mathbf{e}|\mathbf{t})$ and $\boldsymbol{\mu}(\mathbf{t}|\mathbf{e})$ are computed from impostor sets $C(\mathbf{t})$ and $C(\mathbf{e})$. In the following we do not require the cohort sets to be the same as for AS-norm, but we

[1] The original formulation employs an L1-distance, but in practical use-cases L2 distances provide very similar results [15, 17]

Figure 1: *Minimum C^*_{llr} for different normalization approaches, out-of-domain training set (Switchboard + Mixer 04, 05, 06)*



assume that they properly characterize the utterances, so that both (6) and (8) represent the optimal scoring function for the trial. The AS-norm statistics are given by

$$\mu(\mathbf{e}|\mathbf{t}) = \frac{1}{|C(\mathbf{t})|} \sum_{i|\mathbf{x}_i \in C(\mathbf{t})} s(\mathbf{e}, \mathbf{x}_i) = \mathbf{e}^T\mathbf{A}\mathbf{e} + \mathbf{e}^T\mathbf{B}\mathbf{m}_t + \xi_1$$

$$\mu(\mathbf{t}|\mathbf{e}) = \frac{1}{|C(\mathbf{e})|} \sum_{i|\mathbf{x}_i \in C(\mathbf{e})} s(\mathbf{t}, \mathbf{x}_i) = \mathbf{t}^T\mathbf{A}\mathbf{t} + \mathbf{t}^T\mathbf{B}\mathbf{m}_e + \xi_2 \,,$$

where $\mathbf{m}_t$ and $\mathbf{m}_e$ are defined as in (5), and ξ_1, ξ_2 collect all terms that do not depend on $\mathbf{e}$ or $\mathbf{t}$, and thus do not contribute to the intra-trial calibration compensation of score normalization[2]. The normalized score is given by

$$\begin{aligned} s_{asnorm}(\mathbf{e},\mathbf{t}) = &\frac{1}{2}\left(\mathbf{t}^T\mathbf{A}\mathbf{t} + \mathbf{e}^T\mathbf{B}(\mathbf{t}-\mathbf{m}_t)\right) \\ &+ \frac{1}{2}\left(\mathbf{e}^T\mathbf{A}\mathbf{e} + \mathbf{t}^T\mathbf{B}(\mathbf{e}-\mathbf{m}_e)\right) \end{aligned} \tag{10}$$

Comparing (10) with the optimal scoring functions (6) and (8) we can observe that they have similar expressions, however AS-norm is not properly accounting for all terms that appear in the AD-norm expression. The following table compares the terms that appear in either function and depend on the embeddings $\mathbf{e}$ or $\mathbf{t}$:

$s_{adnorm}(\mathbf{e},\mathbf{t})$	$s_{asnorm}(\mathbf{e},\mathbf{t})$
$\mathbf{e}^T\mathbf{A}\mathbf{e}$	$\frac{1}{2}\mathbf{e}^T\mathbf{A}\mathbf{e}$
$-2\mathbf{e}^T\mathbf{A}\mathbf{m}_e$	
$\mathbf{t}^T\mathbf{A}\mathbf{t}$	$\frac{1}{2}\mathbf{t}^T\mathbf{A}\mathbf{t}$
$-2\mathbf{t}^T\mathbf{A}\mathbf{m}_t$	
$\mathbf{e}^T\mathbf{B}\mathbf{t}$	$\mathbf{e}^T\mathbf{B}\mathbf{t}$
$-\mathbf{e}^T\mathbf{B}\mathbf{m}_e$	$-\frac{1}{2}\mathbf{e}^T\mathbf{B}\mathbf{m}_e$
$-\mathbf{t}^T\mathbf{B}\mathbf{m}_t$	$-\frac{1}{2}\mathbf{t}^T\mathbf{B}\mathbf{m}_t$

We observe that the most important term[3] $\mathbf{e}^T\mathbf{B}\mathbf{t}$ is the same in both expressions, whereas the other terms are (mostly) both present, but with a different weight. If we also assume that the dataset shift is not too large, then the contribution of the terms that do not appear in s_{asnorm} but appear in s_{adnorm} or s_{opt} can be considered small. We therefore conclude that for PLDA or PLDA-derived models the mean normalization component of AS-norm can be interpreted as an approximation of the correct scoring function for the given trial. The effectiveness of AS-norm thus suggests that the cohort selection of the latter is providing an effective, *back-end-independent* approach for estimating normalization data that provide a good characterization of the test and enrollment utterances, and motivates our decision to employ the same methodology for computing the cohort sets $C(\mathbf{t})$ and $C(\mathbf{e})$ for AD-norm.

[2]These terms contribute as a global shift of the scores and do not affect the discrimination capabilities of normalized scores.

[3]This term is related to the similarity of the samples, whereas the terms that contain matrix $\mathbf{A}$ are related to the rarity of the trial.

4. Experimental results

To assess the effectiveness of AD-norm we consider a scenario where an off-the-shelf recognizer is employed to evaluate possibly mis-matched trials. The evaluation sets are SRE 2016 [23] and SRE 2019 Evaluation datasets [24]. We consider two different back-ends, PLDA and PSVM, trained on (i) Switchboard and Mixer 04, 05 and 06 data (out-of-domain scenario), or (ii) the same datasets with the addition of the SRE 2018 evaluation set (partially in-domain data for SRE 2019). The normalization sets are the unlabeled portion of SRE 2016 data (2472 segments) and a subset of the SRE 2019 Progress test data (2000 segments), respectively. The embedding extractor is based on the ECAPA architecture [25], trained as in [17]. The 192-dimensional embeddings are normalized by means of Linear Discriminant Analysis, reducing the dimension to 150, and L2-normalization. For PSVM Within Class Covariance Normalization (WCCN) [26] was applied after L2-normalization. In both cases, AD-norm scores are computed by re-centering the embeddings with respect to the mean of the respective cohort set, selected according to the AS-norm criterion as detailed in the previous section. The re-centered embeddings are further L2-normalized. For PSVM, WCCN is applied to the re-centered, L2-normalized embeddings. Experimental results are provided in Figure 1 and Table 1. Figure 1 compares the minimum Cost of Log-Likelihood Ratio [27, 28, 29] C^*_{llr}, which measures the performance of the compared approaches over a wide range of operating points. Table 1 additionally reports Equal Error Rate (EER) and minimum primary cost C^{min}_{prim} as defined by NIST for the two evaluations. We compare our models with five baselines: (i) non-normalized scores, (ii) global mean normalization, computed from the whole impostor set, followed by L2-normalization, (iii) the FDA approach of [5], (iv) AM-norm [7], and (v) AS-norm. For the latter three methods the size of the adaptive cohort set has been set to 200, based on the average results of the better performing baseline, AS-norm. The SRE 2016 results show that all adaptive approaches perform better than global ones regardless of the classification back-end. AM-norm provides results that are very similar to those published by the authors of [7], but performs significantly worse than AS-

Table 1: *Comparison of different normalization approaches on SRE 2016 and SRE 2019 datasets. In bold (i) results of the best* single *method for each training list and back-end combination, and (ii) fusion results equal to or improving those of the best single method.*

		SRE 2016						SRE 2019					
		PLDA			PSVM			PLDA			PSVM		
Training list	Normalization method	EER	C_{prim}^{min}	C_{llr}^{*}	EER	C_{prim}^{min}	C_{llr}^{*}	EER	C_{prim}^{min}	C_{llr}^{*}	EER	C_{prim}^{min}	C_{llr}^{*}
Switchboard + Mixer	None	11.3%	0.97	0.41	12.2%	0.99	0.42	9.7%	0.59	0.34	7.1%	0.59	0.26
	Global	11.0%	0.81	0.38	8.1%	0.74	0.30	9.0%	0.56	0.31	5.6%	0.49	0.21
	FDA [5]	10.5%	0.62	0.36	9.0%	0.58	0.31	9.0%	0.58	0.32	6.9%	0.50	0.25
	AM-norm [7]	9.1%	0.67	0.32	6.9%	0.62	0.25	10.3%	0.67	0.35	6.2%	0.50	0.23
	AS-norm [11]	8.7%	**0.52**	0.30	7.3%	**0.47**	0.26	6.5%	0.52	0.24	**4.9%**	0.43	**0.19**
	AD-norm	**7.6%**	**0.52**	**0.27**	**6.6%**	0.49	**0.23**	**6.2%**	**0.46**	**0.23**	5.1%	**0.41**	**0.19**
	AD+AS-norm	*7.9%*	***0.50***	***0.27***	***6.6%***	***0.46***	***0.23***	***6.1%***	***0.46***	***0.23***	***4.7%***	***0.39***	***0.18***
Switchboard + Mixer + SRE18	None	10.2%	0.87	0.37	8.5%	0.71	0.30	6.4%	0.44	0.23	3.6%	0.36	0.14
	Global	9.7%	0.75	0.35	7.2%	0.60	0.26	6.3%	0.45	0.23	3.8%	0.34	0.15
	FDA [5]	10.3%	0.62	0.36	9.1%	0.58	0.31	8.7%	0.57	0.31	6.8%	0.49	0.25
	AM-norm [7]	8.0%	0.63	0.29	6.5%	0.53	0.24	7.4%	0.51	0.27	4.4%	0.38	0.17
	AS-norm [11]	7.8%	**0.50**	0.27	7.0%	**0.47**	0.25	**4.7%**	0.40	**0.18**	**3.9%**	**0.35**	**0.15**
	AD-norm	**6.9%**	0.52	**0.25**	**6.4%**	0.50	**0.23**	5.0%	**0.38**	0.19	4.1%	0.36	0.16
	AD+AS-norm	*7.2%*	***0.49***	***0.25***	***6.3%***	***0.46***	***0.23***	***4.7%***	***0.37***	***0.18***	***3.8%***	***0.33***	***0.15***

norm in terms C_{prim}^{min}. The two methods achieve similar performance, on average, in terms of EER and C_{llr}^{*}. AD-norm consistently outperforms both approaches in terms of C_{llr}^{*} and EER, although it incurs in a small degradation in terms of C_{prim}^{min} with respect to AS-norm. This suggests that our approach is significantly more effective on a wide range of operating points, although AS-norm performs slightly better in very low false acceptance regions. For SRE 2019 we can observe that AM-norm provides significantly worse results, whereas AS-norm and AD-norm are both competitive. AD-norm is more effective in the mis-matched scenario, whereas AS-norm provides slightly better results for partially matching training and evaluation data. As reference, in-domain PLDA trained on SRE 18 data only achieves an EER of 4.4%, a C_{prim}^{min} of 0.38 and a C_{llr}^{*} of 0.17. Finally, we observe that for PSVM trained with partially matched data none of the considered normalization approaches is able to improve performance with respect to the non-normalized scores, whereas the proposed approach is effective in the mis-matched use-case.

Since AS-norm incorporates a variance normalization effect that has no equivalent in AD-norm, we also evaluated a simple score fusion that computes the average scores of the two models after standardizing non-target scores to unit variance. The results are reported in rows "AD+AS-norm" of Table 1. The fusion proves effective, allowing us to achieve results that are either similar or even slightly better than the best normalization approach, and provides systematic improvements over AS-norm. This motivates us to further investigate possible extensions of AD-norm that incorporate similar effects at embedding rather than score level.

Adaptive approaches require specifying the selected cohort set size K. Figure 2 shows the effects of using different values for K. We can observe that the cohort size has a relevant impact for SRE 2016, whereas the results are more stable on SRE 2019. In practice it may be difficult to select optimal K for a specific dataset without cross-validation data. For some combinations of back-end and normalization model smaller cohort sizes may be more effective, although the relative performance of the approaches remains similar. A cohort size of $K = 200$ provides a good trade-off for the different datasets.

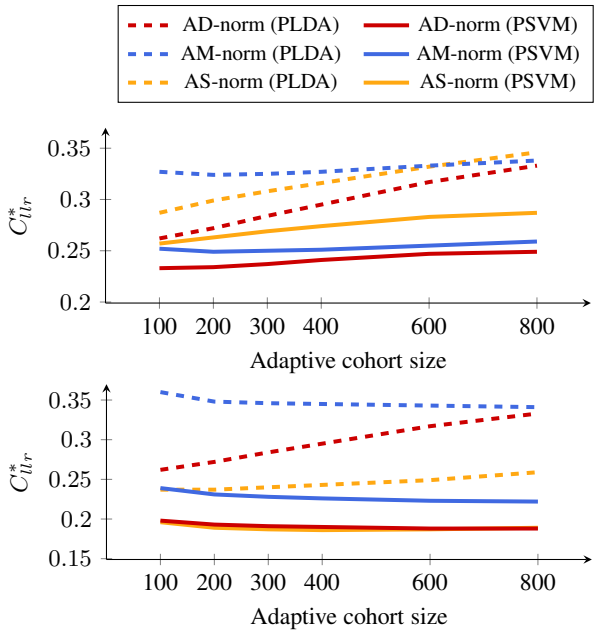


Figure 2: C_{llr}^{*} *as a function of the cohort size. Top: SRE 2016. Bottom: SRE 2019.*

5. Conclusions

We have presented a framework for the normalization of speaker embeddings that employs the AS-norm cohort selection mechanism to select effective normalization cohorts. These cohort sets are then used to estimate an utterance-dependent mean vector that is employed to re-center the enrollment and test speaker embeddings. The proposed AD-norm allows improving the accuracy of a speaker recognition system in case of severe mismatch between training and evaluation data, providing similar or better performance with respect to AS-norm without the AS-norm scoring-time computational costs.

6. References

[1] T. Pekhovsky and A. Sizov, "Comparison between supervised and unsupervised learning of Probabilistic Linear Discriminant Analysis Mixture models for speaker verification," *Pattern Recognition Letters*, vol. 34, pp. 1307–1313, 2013.

[2] D. Garcia-Romero, A. McCree, S. Shum, N. Brümmer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proc. of Odyssey 2014, The Speaker and Language Recognition Workshop*, 2014, pp. 260–264.

[3] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4002–4006.

[4] K. A. Lee, Q. Wang, and T. Koshinaka, "The coral+ algorithm for unsupervised domain adaptation of plda," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5821–5825.

[5] P.-M. Bousquet and M. Rouvier, "On Robustness of Unsupervised Domain Adaptation for Speaker Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2958–2962.

[6] R. Li, W. Zhang, and D. Chen, "The coral++ algorithm for unsupervised domain adaptation of speaker recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7172–7176.

[7] M. Mclaren, M. H. Rahman, D. Castan, M. K. Nandwana, and A. Lawson, "Adaptive Mean Normalization for Unsupervised Adaptation of Speaker Embeddings," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 88–94.

[8] 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.

[9] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, p. 42–54, jan 2000.

[10] D. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker recognition," in *Proc. of ICASSP*, 2005.

[11] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Proceedings of Interspeech 2011*, 2011, pp. 2365–2368.

[12] Z. Karam, W. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *Proc. of ICASSP*, 05 2011, pp. 4512–4515.

[13] A. Swart and N. Brümmer, "A Generative Model for Score Normalization in Speaker Recognition," in *Proc. Interspeech 2017*, 2017, pp. 1477–1481.

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

[15] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Černocký, "Analysis of score normalization in multilingual speaker recognition," in *Proceedings of Interspeech*, 2017.

[16] 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 *Proceedings of Interspeech*, 2017.

[17] S. Cumani and S. Sarni, "Impostor Score Statistics as Quality Measures for the Calibration of Speaker Verification Systems," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 25–32.

[18] 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.

[19] 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.

[20] 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.

[21] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text–independent speaker verification," in *Proceedings of Odyssey 2010*, 2010, pp. 76–82.

[22] S. Cumani and P. Laface, "Exact memory–constrained upgma for large scale speaker clustering," *Pattern Recognition*, vol. 95, pp. 235–246, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319302493

[23] "The NIST 2016 speaker recognition evaluation plan," 2016, available at "https://www.nist.gov/system/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf".

[24] "The NIST 2019 speaker recognition evaluation: Cts challenge," 2019, available at http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf.

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

[26] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proceedings of ICSLP 2006*, 2006, pp. 1471–1474.

[27] 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.

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

[29] 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.