Deep Quality-informed Score Normalization for Privacy-friendly SpeakerRecognition in unconstrained Environments

Andreas Nautsch[1], Søren Trads Steen[1,2], Christoph Busch[1]

Abstract: In scenarios that are ambitious to protect sensitive data in compliance with privacy regulations, conventional score normalization utilizing large proportions of speaker cohort data is not feasible for existing technology, since the entire cohort data would need to be stored on each mobile device. Hence, in this work we motivate score normalization utilizing deep neural networks. Considering unconstrained environments, a quality-informed scheme is proposed, normalizing scores depending on sample quality estimates in terms of completeness and signal degradation by noise. Utilizing the conventional PLDA score, comparison i-vectors, and corresponding quality vectors, we aim at mimicking cohort based score normalization optimizing the $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ discrimination criterion.

Examining the I4U data sets for the 2012 NIST SRE, an 8.7% relative gain is yielded in a pooled 55-condition scenario with a corresponding condition-averaged relative gain of 6.2% in terms of $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$. Robustness analyses towards sensitivity regarding unseen conditions are conducted, i.e. when conditions comprising lower quality samples are not available during training.

Keywords: speaker recognition, score normalization, unconstrained environments, neural networks, deep learning

1 Introduction

Accounting for European data privacy regulations [Eu16], resource limitations of mobile operating scenarios, and technological requirements concerning vast signal quality variations in unconstrained environment speaker recognition, current score normalization schemes are put to its limits. In this paper, we propose a quality-informed score normalization scheme utilizing cohort data for the purpose of training a neural network in order to avoid a distribution of biometric data from cohort subjects, substituting conventional cohort-based score normalization. This study is limited with respect to deeper network architectures and the sensitivity to unseen quality conditions. Comparative experiments to conventional normalization schemes are excluded, since we assume their design to be prohibited due to a restrictive interpretation of §9 in EU regulation 2016/679, i.e. cohort data which is necessary to estimate parameters of zero-norms shall not be distributed. The EU regulation 2016/679 [Eu16, §9] prohibits the processing of biometric data, if not – among others – the biometric subject is giving consent, and the *processing relates to personal data which are manifestly made public by the data subject*. Hence, the distribution and use of cohort data related to other individuals than the biometric subject under processing may

[1] da/sec — Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {andreas.nautsch,christoph.busch}@{crisp-da | h-da}.de
[2] Technical University of Denmark, Denmark, stradssteen@gmail.com

become improper to justify as cohort data would need to be transmitted to the device of any other biometric user for conducting cohort normalization, especially for data deletion.

This paper is organized as follows: Sec. 2 depicts the related work on speaker recognition and neural networks. Sec. 3 depicts the proposed normalization scheme. Experimental evaluations are carried out in Sec. 4, and conclusions are drawn in Sec. 5.

2 Related Work

Recent speaker recognition approaches rely on i-vectors, representing the characteristic speaker offset from an Universal Background Model (UBM), which models the distribution of acoustic features, such as Mel-frequency cepstral coefficients [RQD00]. Thereby, UBM components' mean vectors are concatenated to a *supervector* $\vec{\mu}_{\text{UBM}}$. Speaker supervectors $\vec{s}$ are decomposed by a total variability matrix into a lower-dimensional i-vector $\vec{i}$ as an offset to the UBM supervector $\vec{\mu}_{\text{UBM}}$ [Ke05, De11]. Then, i-vectors are projected onto a spherical space by whitening transform and length normalization [GREW11, BBM13]. State-of-the-art i-vector comparators, e.g. Probabilistic Linear Discriminant Analysis (PLDA) [CL14], conduct a likelihood ratio scoring.

2.1 Conventional Score Normalization Methods

State-of-the-art recognition systems [Va16, Br16] utilize score normalization in order to improve discrimination power on secure operating points by employing statistics from comparisons of the reference against an independent (cohort) data set, referred to as z-norm, from comparisons of the probe against a cohort set, referred to as t-norm, and variations of z- and t-norm, such as the zt-norm, or s-norm, as well as adaptive variations e.g., at- [SR05] and as-norm [Cu11]. Exemplary, in [SR05, Ha13], data of 550, 1039 female, and 435, 680 male speakers is utilized for normalization purposes, respectively, whereas in [Cu11], solely the usage of 348 female and 273 male voice samples is reported. In mobile applications, where no data of the biometric subject should leave the device, the cohort data needs to be present on each mobile device.

2.2 Different Environmental Conditions

Variations in signal quality, i.e. in the probe sample condition, result in different score distributions per condition [Ma13, MSvL15]. While systems are usually calibrated for known scenarios and in fixed-condition environments, calibrating systems well among known as well as unseen conditions is harder, i.e. when facing unconstrained environments.

In this paper, we examine the 55 duration and noise conditions presented in [Na15]. In [Na15], SNR conditions stem from two noise sources: air conditioner (AC) and crowd (CROWD) noise. By degenerating voice samples from the I4U file list [Sa13], combined signal degradation and observation incompleteness (short probe segment duration) effects are simulated, which are expected to represent the most common conditions, cf. Tab. 1.

Tab. 1: *Label scheme for combined duration and noise conditions, cf. [Na15].*

Condition	1	2	3	4	5	6	7	8	9	10	11 … 30	31 … 55
Duration	5 s	10 s	20 s	40 s	full	5 s					10 s … full	5 s … full
Noise	clean					AC						CROWD
SNR						0 dB	5 dB	10 dB	15 dB	20 dB	0 dB … 20 dB	0 dB … 20 dB

2.3 Estimation of Unified Audio Quality Vectors

For the purpose of estimating quality in speaker recognition, unified audio characteristics [Fe12] are utilized. Single multivariate Gaussian models $\Lambda_j \sim \mathcal{N}(\mu_j, \Sigma), j = 1, \ldots, 55$ are trained in original i-vector space for each quality condition as outlined in Tab. 1. The models have condition-dependent mean vectors μ_j and share a full covariance matrix Σ. Class-dependent means are estimated using i-vectors from a respective quality condition and Σ is estimated by pooling all the i-vectors. The resulting vector of posterior probabilities for an i-vector $\vec{i}$ represent a condition quality vector (q-vector) $\vec{q}$ [Fe12], with entries:

$$q(j) = \frac{P(\vec{i}\,|\,\Lambda_j)}{\sum_{j=1}^{55} P(\vec{i}\,|\,\Lambda_j)}. \tag{1}$$

2.4 Neural Network schemes

Feed forward neural networks consist of layers of units [Bi06]. An input layer and an output layer are linked over a number of hidden layers by numerous connections, where the connections between units of each layer are weighted. In [He15], initial weights are proposed having a standard deviation of $\sqrt{2/n_l}$, with n_l being the number of incoming connections to the unit. In each unit, a linear combination, the *response*, is constructed from the outputs of the previous layer's units. A non-linear activation function is evaluated on the response to achieve the output, or *activation*, of the units e.g., the *linear rectifier, ReLU* activation function [LBH15] and the sigmoid function for bounded activations [Bi06]. Networks are trained to optimize the performance regarding the cost function using gradient descent, where the Adam algorithm [KB14] and *backpropagation* [Bi06] can be employed. As a cost function, the binary cross-entropy function is a measure of the distance between the distribution of the actual classes and the distribution of the prediction. In this work, we utilize a single-unit output layer, representing a system's score. In order to avoid over-fitting of the training data, different regularization schemes can be employed, such as *weight decay* [Bi06], *dropout* [Sr14], and *batch normalization* [IS15].

3 Deep Quality-informed Normalization

In order to account for cohort-related data as well as quality information, we propose to construct the input layer to a feed forward neural network based on the comparison score, reference and probe i-vectors $\vec{i}_{ref}, \vec{i}_{prb}$ as well as corresponding q-vectors $\vec{q}_{ref}, \vec{q}_{prb}$, cf. Fig. 1, whereas a normalized score between 0 and 1, representing impostor and genuine classes, respectively, is obtained via a single unit output layer with a sigmoid activation function, yielding rather discriminative than well-calibrated scores. By training the network on the cohort data set, we assume the network model to comprise cohort and quality information, whilst achieving anonymity (not only pseudonymization) for the cohort speakers. Furthermore, massive data amounts featuring multi-condition quality is not required to be transferred to each mobile device by the biometric system operator.

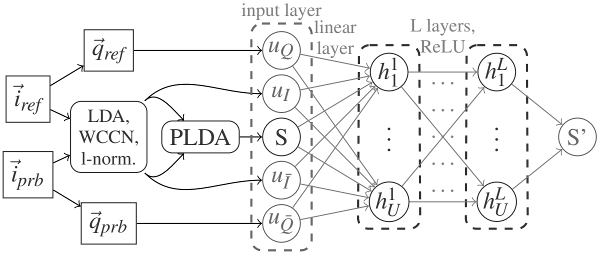


Fig. 1: Proposed deep quality-informed normalization network design with input layer (green), hidden layers (red), and output layer score S'. $u_Q, u_{\bar{Q}}$ represent Q-dim. q-vectors, and $u_I, u_{\bar{I}}$ represent I-dim. i-vectors, respectively.

For the purpose of accounting for linear normalization approaches e.g., linear quality calibration [Bd11, Fe12, Na16], the first hidden layer of the proposed network employs a linear activation function $f(x) = a + bx$. During training, input features are adaptively normalized with respect to the amount of genuine and impostor comparisons. Deeper hidden layers are non-linear using the ReLU activation function. The weights are initialized by the scheme proposed in [He15]. Convergence is reached after 3 epochs on a random-selected 20% held-out validation subset, on which the best performing model is chosen. In order to achieve an effective class balance of equal priors, genuine comparisons are weighted higher than the impostor comparisons during network training. The network configuration is referred to as (L, U) with a network of a linear layer with U units, followed by L non-linear layers of U units, cf. Fig. 1.

4 Experimental Set-Up and Analysis

For the purpose of studying the proposed method, first we examine regularization impacts on a fixed configuration of number of layers and units, finding $\lambda = 10^{-5}$ to reduce overfitting well, then parameters of the deep neural network with fix regularization parameter are examined comparing reasonable configurations on the testing set. In order to gain insights on the robustness of the proposed normalization scheme, a sensitivity analysis is conducted by excluding poor quality conditions from training the normalization network.

Implementations are based on Python 2.7 with Keras 1.1.1 and Theano 0.9.0.dev1, Matlab 2016b, and the BOSARIS toolkit [Bd11]. The data used is the same as in [Na15] of the I4U file list for NIST SRE'12 [Sa13]. The dataset consists of 55 different degradations in duration and noise type and level, denoted here as degradation conditions, cf.

Tab. 1. There are 680 reference i-vectors and 357269 probe i-vectors in the training dataset, and 723 reference and 388278 probe i-vectors in the test set. The i-vectors are processed dependent on the noise condition by performing linear discriminant analysis (LDA) to 200 dimensions, within class covariance normalization (WCCN), and length normalization [GREW11]. Baseline scores are derived from our recent work [Na15, Na16].

As an application-independent performance metric, we use minimum cost of log-likelihood ratio scores $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ [BdP08], i.e. the generalized empirical cross-entropy of genuine and impostor scores, assuming well-calibrated systems in terms of Bayes decisions. The upper bound of $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ is determined by the EER of the ROC's convex hull [Bd11].

4.1 Experimental Analysis: Network Configuration

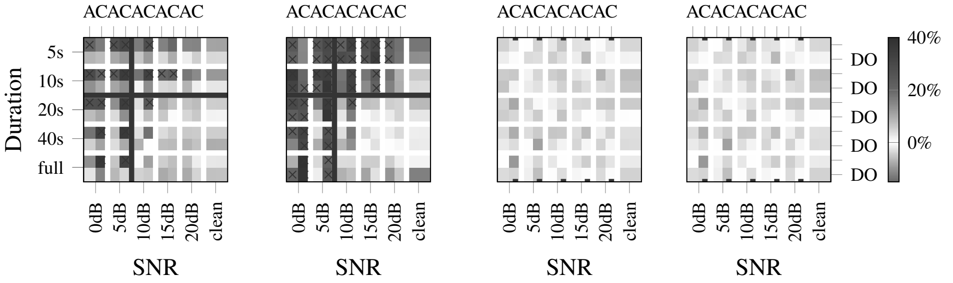
In order to examine network configurations, we investigate on $L = 1,2,4$ layers, where all layers comprise the same amount of hidden units, i.e. $U = 50,100,200$ units. Tab. 2 compares the different networks on the test set: configuration (1, 50) yields the largest condition-average $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ gain over a conventional i-vector / PLDA baseline system of 6.2% with the lowest standard variation, i.e. with rather stable improvements among all conditions. Configuration (2, 100) yields the second largest gains regarding average and deviation in terms of $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$, but also regarding pooled-condition performance, where the (2, 50) network yields the largest gains. Accounting for potential over-fitting, dropout is examined on (1, 50) and (2, 100) networks with a 20% dropout rate: on average, $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ grows, which may occur due to a too high dropout rate. Further investigations are carried out on the (1, 50) configuration, due its gains on pooled performance.

Tab. 2: Benchmark of relative $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ changes (in %) to PLDA baseline on the test set regarding condition averaging (μ), standard deviation (σ), and pooling (p), and dropout training (DO).

(L, U)	(1, 50)	(1, 100)	(1, 200)	(2, 50)	(2, 100)	(2, 200)	(4, 50)	(4, 100)	(4, 200)	(1, 50)	(2, 100)
										DO	
μ	**-6.2**	1.4	-2.0	-2.1	***-5.7***	0.8	-2.9	-5.2	-0.9	*1.3*	4.9
σ	**2.4**	6.6	3.5	4.3	***2.6***	4.4	3.1	2.9	3.8	*1.6*	4.0
p	-4.6	0.9	-0.2	**-6.6**	***-6.4***	0.4	-0.2	-3.4	0.0	*-2.5*	7.1

4.2 Robustness Analysis to unseen signal degradation and noise types

For the purpose of examining the robustness of the proposed normalization, training is conducted with unseen test conditions, i.e. all conditions afflicted with SNR levels $\leq 5\ dB$ and with durations $\leq 10\ s$ are excluded. Figs. 2a, 2b compare the effects to (1, 50) and (2, 100) configurations, with and without employing dropout, regarding whether or not the $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$ performance is not exceeding a ±20% performance band with respect to each condition's $\mathrm{C}_{\mathrm{llr}}^{\mathrm{min}}$. In this analysis, the (1, 50) configuration outperforms the (2, 100) in terms of coherence stability. Also, employing dropouts sustain coherent and stable performance. By placing focus on robustness towards noise type rather than low-SNRs, we exclude all CROWD noise afflicted conditions from training instead: both configurations perform stable and coherent with slight benefits from conducting dropout training, see Figs. 2c, 2d.



(a) conf. (1, 50), low-SNR and short duration conditions excluded

(b) conf. (2, 100), low-SNR and short duration conditions excluded

(c) conf. (1, 50), CROWD noise conditions excluded

(d) conf. (2, 100), CROWD noise conditions excluded

Fig. 2: Relative $C_{\text{llr}}^{\text{min}}$ change on test set (in %). Performance by duration and SNR regarding AC (A) and CROWD (C) noise as well as whether dropout is conducted (DO). Green lines indicate the conditions excluded from training. Crosses denote relative $C_{\text{llr}}^{\text{min}}$ changes above ±20%.

4.3 Summary and Discussion

Examining deeper architectures considering non-linear layers, gains compared to the baseline PLDA performance are observed on average, though not further increasing the single linear layer performance. Comparatively, the cohort normalization in [Na15] yields up to 8.2% relative gains in $C_{\text{llr}}^{\text{min}}$ on single conditions. In the robustness analysis, i.e. by excluding poor quality conditions and the more challenging noise type, the proposed approach reveals to benefit on good quality conditions, the performance of the (1, 50) configuration is preserved within a ±20% performance band on unseen poor quality conditions. Contrastively, on excluding overlapping speech (CROWD noise) conditions, either (1, 50) and (2, 100) configurations perform comparatively stable. Thus, the proposed approach benefits rather from training on a broad scale of SNR levels than on more noise types, posing a challenging scenario due to overlapping biometric features of other subjects.

5 Conclusion

In this study, we introduced a neural network based normalization approach utilizing quality estimates, suitable for unconstrained environments under data privacy as well as limited resource concerns regarding the data of cohort speakers. As system operators transmit trained networks to mobile devices instead of cohort data, data privacy is achieved for cohort subjects, while sustaining comparative discrimination performance. Robustness analyses show benefits of knowing levels of SNR levels and durations during training over knowing different noise types of mid / high-SNR levls during training.

6 Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts within the Center for Research in Security and Privacy (`www.crisp-da.de`), and the BioMobile II project (no. 518/16-30).

References

[BBM13] Bousquet, P.-M.; Bonastre, J.-F.; Matrouf, D.: Identify the Benefits of the Different Steps in an i-Vector Based Speaker. Springer-Verlag Berlin Heidelberg, chapter CIARP, Part II, pp. 278–285, 2013.

[Bd11] Brümmer, N.; de Villiers, E.: The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing. Technical report, AGNITIO Research, South Africa, December 2011. Last accessed: 2017-05-15.

[BdP08] Brümmer, N.; du Preez, J.: Application-Independent Evaluation of Speaker Detection. Computer Speech and Language, 20(2):230–275, July 2008.

[Bi06] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC, 2006.

[Br16] Brümmer, N.; Swart, A.; Jorrím-Prieto, J.; García, P. et al.: ABC NIST SRE 2016 System Description. In: Proc. of the NIST SRE 2016 workshop. 2016.

[CL14] Cumani, S.; Laface, P.: Generative pairwise models for Speaker Recognition. In: Proc. Odyssey 2014: The Speaker and Language Recognition Workshop. 2014.

[Cu11] Cumani, S.; Batzu, P. Domenico; Colibro, D.; Vair, C.; Laface, P.; Vasilakakis, V.: Comparison of Speaker Recognition Approaches for Real Applications. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). 2011.

[De11] Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transaction on Audio, Speech, and Language Processing (TASLP), 19(4):788–798, May 2011.

[Eu16] European Council: , Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), April 2016.

[Fe12] Ferrer, L.; Burget, L.; Plchot, O.; Scheffer, N.: A Unified Approach for Audio Characterization and its Application to Speaker Recognition. In: Odyssey 2012: The Speaker and Language Recognition Workshop. 2012.

[GREW11] Garcia-Romero, D.; Epsy-Wilson, C.Y.: Analysis of i-vector length normalization in Speaker Recognition systems. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). pp. 249–252, 2011.

[Ha13] Hasan, T.; Saeidi, R.; Hansen, J. H. L.; van Leeuwen, D. A.: Duration Mismatch Compensation for i-vector based Speaker Recognition systems. In: Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2013.

[He15] He, K.; Zhang, X.; Ren, S.; Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE Intl. Conf. on Computer Vision. pp. 1026–1034, 2015.

[IS15] Ioffe, S.; Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[KB14] Kingma, D.; Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[Ke05] Kenny, P.: Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. Technical Report CRIM-06/08-13, CRIM, Montreal, 2005.

[LBH15] LeCun, Y.; Bengio, Y.; Hinton, G.: Deep learning. Nature, May 2015.

[Ma13] Mandasari, M. I.; Saeidi, R.; McLaren, M.; van Leeuwen, D. A.: Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions. IEEE Trans. on Audio, Speech and Language Processing (TASLP), 21(11):2425–2438, 2013.

[MSvL15] Mandasari, M. I.; Saeidi, R.; van Leeuwen, D. A.: Quality measures based Calibration with Duration and noise dependency for Speaker Recognition. Speech Communication, 72:126–137, September 2015.

[Na15] Nautsch, A.; Saeidi, R.; Rathgeb, C.; Busch, C.: Analysis of mutual Duration and noise effects in Speaker Recognition: benefits of condition-matched cohort selection in Score Normalization. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). pp. 3006–3010, 2015.

[Na16] Nautsch, A.; Saeidi, R.; Rathgeb, C.; Busch, C.: Robustness of Quality-based Score Calibration of Speaker Recognition Systems with respect to low-SNR and short-Duration conditions. In: Proc. of Odyssey 2016: The Speaker and Language Recognition Workshop. pp. 358–365, 2016.

[RQD00] Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Conversational Speech, Digital Signal Processing, 10:19–41, 2000.

[Sa13] Saeidi, R.; Lee, K.A.; Kinnunen, T. et al.: I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). ISCA, 2013.

[SR05] Sturim, D.E.; Reynolds, D.A.: Speaker adaptive Cohort Selection for tnorm in text-dependent Speaker Verification. In: Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2005.

[Sr14] Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1):1929–1958, 2014.

[Va16] Vair, C.; Colibro, D.; Dalmasso, E.; Farrell, K. et al.: Nuance - Politecnico di Torino (NPT) System Description for NIST 2016 Speaker Recognition Evaluation. In: Proc. of the NIST SRE 2016 workshop. 2016.