INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Analysis of Score Normalization in Multilingual Speaker Recognition

Pavel Matějka, Ondřej Novotný, Oldřich Plchot, Lukáš Burget, Mireia Diez Sánchez, Jan "Honza" Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia

{matejkap,inovoton,iplchot,burget,mireia,cernocky}@fit.vutbr.cz

Abstract

NIST Speaker Recognition Evaluation 2016 has revealed the importance of score normalization for mismatched data conditions. This paper analyzes several score normalization techniques for test conditions with multiple languages. The best performing one for a PLDA classifier is an adaptive s-norm with 30% relative improvement over the system without any score normalization. The analysis shows that the adaptive score normalization (using top scoring files per trial) selects cohorts that in 68% contain recordings from the same language and in 92% of the same gender as the enrollment and test recordings. Our results suggest that the data to select score normalization cohorts should be a pool of several languages and channels and if possible, its subset should contain data from the target domain.
Index Terms: speaker recognition, score normalization

1. Introduction

For speaker verification systems, score normalization is one of the standard steps in producing well calibrated speaker verification scores. Without the normalization, different distributions of target and non-target scores[1] can be obtained for two different enrolled speaker models. This makes it impossible to set a single detection threshold for the scores obtained from the different speaker models. Similarly, for the same speaker model, the score distributions can vary depending on the test utterance condition (recording channel, acoustic conditions, language of the utterance, etc.) which calls for a condition dependent threshold. Already the early works on automatic speaker recognition by Reynolds [1, 2, 3] reported degraded performance on the mismatched conditions and stressed the importance of score normalization.

Typically, the normalization step shifts and scales the distributions for the individual models and/or conditions to allow for a single detection threshold. The shifts and scales are usually estimated using a set of utterances so called normalization cohort. It has been shown many times that for GMM-UBM algorithms [4, 1] and later especially for Joint Factor Analysis (JFA) based system [5, 6], we might achieve significantly better accuracy with score normalization, such as Z-norm [4], T-norm [7], combinations of both (TZ-norm and ZT-norm) [8] or other variants such as H-norm [1], D-norm [9], KL-T-norm [10], S-norm [11], normalized cosine similarity [12], speaker clusters [13] and many others [14, 15, 16].

Adaptive T-norm [17], Top-norm [18] or Adaptive S-norm [19, 20] are modifications of the basic techniques selecting not all speakers from the normalization cohort but only the top-scoring ones. This approach can be actually applied to almost any score normalization and is denoted *adaptive score normalization* in the rest of our paper.

Based on the analysis reported in [21], score normalization was the key element for the best performance in the NIST Speaker recognition evaluation 2016, despite the fact that the best systems in NIST SRE 2010 and 2012 did not perform the score normalization. We believe that the main reason for this is that both 2010 and 2012 evaluations were English only, with channels matching the previous NIST evaluations: therefore, there was plenty of data for matched-condition PLDA training, which does not require score normalization. On the other hand, score normalization has a big effect in mismatched conditions, as it can shift and scale the scores by examining how the system behaves on the new data.

This paper compares several normalization techniques as well as different cohorts and analyzes the nature of files selected to the cohort in adaptive score normalization. The experiments are done on several conditions.

2. Score normalization techniques

The goal of score normalization is to reduce within trial variability leading to improved performance, better calibration, and more reliable threshold setting. This section describes the most used score normalization techniques. Below, the score between enrollment utterance e and test utterance t is denoted $s(e,t)$.

2.1. Z-norm

Zero score normalization [4] employs impostor score distribution for enrollment file. It uses a cohort $\mathcal{E} = \{\varepsilon_i\}_{i=1}^{N}$ with N speakers which we assume to be different from the speakers in utterances e and t. The cohort scores are

$$S_e = \{s(e,\varepsilon_i)\}_{i=1}^{N} \tag{1}$$

and are formed by scoring enrollment utterance e with all files from cohort $\mathcal{E}$. The normalized score is then:

$$s(e,t)_{z\text{-}norm} = \frac{s(e,t) - \mu(S_e)}{\sigma(S_e)}, \tag{2}$$

where $\mu(S_e)$ and $\sigma(S_e)$ are mean and standard deviation of S_e.

2.2. T-norm

Test score normalization [7] is similar to Z-norm with the difference that it normalizes the impostor score distribution for the test utterance. T-norm can be expressed by:

This work was supported by the DARPA RATS Program un- der Contract No. HR0011-15-C-0038. The views expressed are those of the author and do not reflect the official policy or posi- tion of the Department of Defense or the U.S. Government. The work was supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", European Union's Horizon 2020 project No. 645523 BISON, by Google research award, Grant Agency of the Czech Republic project No. GJ17-23870Y, by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602" and European Union's Horizon 2020 Marie Sklodowska-Curie grant agreement No. 748097.

[1]Target and non-target scores are obtained by scoring enrolled speaker model w.r.t. target and imposter test utterances, respectively.

http://dx.doi.org/10.21437/Interspeech.2017-803

$$S_t = \{s(t, \varepsilon_i)\}_{i=1}^{N} \quad (3)$$

$$s(e,t)_{t\text{-}norm} = \frac{s(e,t) - \mu(S_t)}{\sigma(S_t)} \quad (4)$$

where $\mu(S_t)$ and $\sigma(S_t)$ are mean and standard deviation of S_t.

2.3. ZT-norm

ZT-norm or TZ-norm use Z- and T-norm in series, and might use different cohorts for each step [22]. By doing this, the scores are normalized with respect to both enrollment and test utterances. Applying ZT-norm for Joint Factor Analysis (JFA) based system was an essential step, which improved results by 50% relative [5, 6].

2.4. S-norm

The symmetric normalization (S-norm) computes an average of normalized scores from Z-norm and T-norm [11]. S-norm is symmetrical as $s(e,t) = s(t,e)$, while the previously mentioned normalizations depend on the order of e and t.

$$\begin{aligned} s(e,t)_{s\text{-}norm} &= \frac{1}{2} \cdot (s(e,t)_{z\text{-}norm} + s(e,t)_{t\text{-}norm}) \\ &= \frac{1}{2} \cdot \left(\frac{s(e,t) - \mu(S_e)}{\sigma(S_e)} + \frac{s(e,t) - \mu(S_t)}{\sigma(S_t)} \right) \end{aligned} \quad (5)$$

2.5. Adaptive score normalization

In adaptive T-norm [17] or Top-norm [18], only part of the cohort is selected[2] to compute mean and variance for normalization. We investigated the same cohort selection for Z-norm, T-norm, ZT-norm and S-norm - we call this selection adaptive, as the selected cohort might change for every speaker.

Two variants of adaptive cohort selection can be found in the literature: the adaptive cohort can be either selected to be X closest (most positive scores) files to the enrollment file $\mathcal{E}_e^{top}$, or, as in [20], to the test file $\mathcal{E}_t^{top}$. We have to note that such cohorts are different for each enrollment utterance e or test utterance t respectively. The cohort scores based on such selections for the enrollment utterance are then:

$$S_e(\mathcal{E}_e^{top}) = \{s(e,\varepsilon)\}_{\forall \varepsilon \in \mathcal{E}_e^{top}}, \quad S_e(\mathcal{E}_t^{top}) = \{s(e,\varepsilon)\}_{\forall \varepsilon \in \mathcal{E}_t^{top}} \quad (6)$$

and correspondingly for the test utterance t.

Two variants were investigated with S-norm: the normalized score for the first one called **adaptive S-norm1** is

$$\begin{aligned} s(e,t)_{as\text{-}norm1} = \frac{1}{2} \cdot \Bigg(&\frac{s(e,t) - \mu(S_e(\mathcal{E}_e^{top}))}{\sigma(S_e(\mathcal{E}_e^{top}))} + \\ &+ \frac{s(e,t) - \mu(S_t(\mathcal{E}_t^{top}))}{\sigma(S_t(\mathcal{E}_t^{top}))} \Bigg) \end{aligned} \quad (7)$$

and the second variant, **adaptive S-norm2**, is defined as

$$\begin{aligned} s(e,t)_{as\text{-}norm2} = \frac{1}{2} \cdot \Bigg(&\frac{s(e,t) - \mu(S_e(\mathcal{E}_t^{top}))}{\sigma(S_e(\mathcal{E}_t^{top}))} + \\ &+ \frac{s(e,t) - \mu(S_t(\mathcal{E}_e^{top}))}{\sigma(S_t(\mathcal{E}_e^{top}))} \Bigg) \end{aligned} \quad (8)$$

Adaptive S-norm2 was successfully used by Nuance in the NIST SRE 2016 evaluation [23].

[2] Usually X top scoring or most similar files, where X is set to be for example 200

3. Experimental setup

3.1. Datasets

Evaluation sets

- *sre16evl* – NIST 2016 Speaker Recognition Evaluation data is our main evaluation set because it contains new languages and channels not represented in the training data. The dataset comprises utterances in two languages, Tagalog and Cantonese. Enrollment files have nominal durations of 60 s of speech whereas the durations of test files range from 10 to 60 s. The set is composed of 37058 target and 1949462 nontarget trials for the pooled (female + male) condition. For more details see [24].
- *sre10c05* – NIST 2010 Speaker Recognition Evaluation data for telephone-telephone condition [25]. This dataset is a well known one and the majority of published papers report results on it since 2010. Previous NIST evaluations contain lots of data for matched training. It contains only English utterances with nominal durations of around 120 s of speech. It consists of 3704 target and 233077 nontarget female trials and 3465 target and 175873 nontarget male trials.
- *lan-lan* refers to the language-language condition defined in the PRISM set [26], which comprises data from previous NIST evaluations in five different languges. The distribution of languages is displayed in Table 3 (numbers in brackets). In addition to original PRISM set files, we generated short cuts from these files, resembling the histogram of durations of the sre16evl set. The motivation for evaluating on this set is that it contains multiple languages and similar channel to the training data. In total, it contains 37503 target and 796449 nontarget female trials and 20308 target and 347484 nontarget male trials.

Normalization Cohorts

- *NIST* – contains one utterance per speaker from NIST SRE training data. It comes from different channels and contains only limited amount of data with the same languages as sre16evl. We experimented with different selection methods not to include many utterances of each speaker to the cohort, but there was no significant change in the results.
- *LID* – contains files in many languages from our development and evaluation data from the past NIST LRE evaluations [27, 28]. This set should better address the language mismatch problem. In total, there are 75k files from 57 languages with nominal durations longer than 30 s of speech.
- *SRE16 unlabeled* – provided as development data for NIST SRE 2016 [24] as matched channel and language data. This set should be ideal for score normalization for sre16evl. There are 200 files from minor languages (Cebuano and Mandarin) and 2272 files for major (matched) languages (Tagalog and Cantonese).
- *SRE16 minor* – this set mimics the scenario of similar recording channels but different languages. It contains labeled development data for NIST SRE 2016 (20 speakers, each with 10 calls in minor languages) and 200 files with minor languages from the "SRE16 unlabeled" set.

3.2. Evaluation metrics

Results are reported as in terms of $\mathrm{DCF}^{\min}$ as defined for NIST SRE 2016. The DCF value is obtained as an average of two operating points with P_{tar}=0.01 and P_{tar}=0.005, see the evaluation plan [24] for more details.

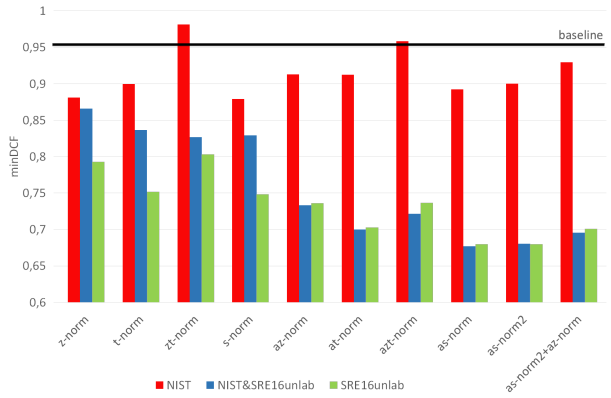

Figure 1: *Comparison of different score normalization techniques -* $\mathrm{DCF}^{\mathrm{min}}$ *on all pooled trials from NIST SRE 2016*

Table 1: $\mathrm{DCF}^{\mathrm{min}}$ *for different score normalization techniques for pooled NIST SRE 2016 with different datasets in cohort.*

norm. / cohort	NIST	NIST& SRE16unlab	SRE16unlab
baseline (no-norm)	0,9539	0,9538	0,9538
z-norm	0,8811	0,8661	0,7926
t-norm	0,8996	0,8366	0,7514
zt-norm	0,9814	0,8270	0,8029
s-norm	**0,8790**	0,8294	0,7483
az-norm1	0,9131	0,7335	0,7362
at-norm1	0,9124	0,6998	0,7026
azt-norm1	0,9584	0,7214	0,7365
as-norm1	0,8922	**0,6771**	**0,6797**
as-norm2	0,8999	0,6806	**0,6797**
as-norm2+az	0,9293	0,6954	0,7010

3.3. System description

Our system employs gender independent i-vector extraction and PLDA scoring [29, 11, 30]. The front-end operates on standard 19 Mel-Frequency Cepstrum Coefficients (MFCC) with C0, delta and double deltas, which are short-term mean and variance normalized over a 3 s sliding window. The universal background model (UBM) has 2048 diagonal-covariance Gaussians and the i-vector extractor produces 600-dimensional vectors. UBM was trained on approximately 8500 telephone files (313 hours of speech after VAD), i-vector extractor on 75000 files (3650 hours) and PLDA on 121000 files (5300 hours) defined by PRISM set [26]. Additionally, we generated utterances with artificially added noise, reverberation and short cuts from non-English files which were added to PLDA training to simulate the properties of the data in NIST SRE 2016[3].

4. Results

4.1. Comparison of score normalization

This section presents primarily results on NIST SRE 2016, as it introduced new variabilities which were not present in previous evaluations. Later, we complete the analysis on the well known NIST SRE 2010 telephone–telephone (c5) condition and also on the "other language" conditions from PRISM set.

Figure 1 and Table 1 show the results on the all pooled NIST SRE16 trials with different score normalization techniques and

[3]The number of files and hours of speech above are already with these additional noisy and reverberated files.

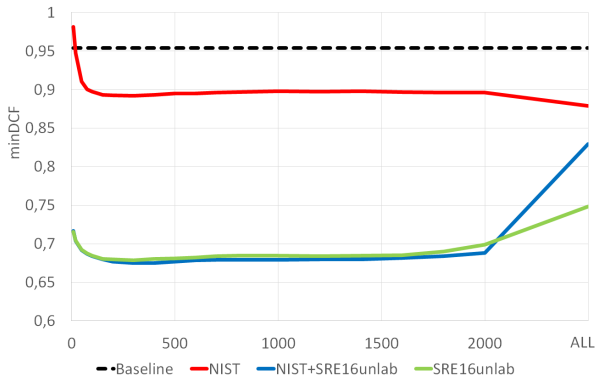

Figure 2: *Numbers of files selected to the cohort in adaptive S-norm - results are in* $\mathrm{DCF}^{\mathrm{min}}$ *on all pooled trials from NIST SRE 2016*

with 3 different cohort sets for normalization. NIST cohort set contain different languages and different channel than evaluation data, SRE16unlab contains the same languages and channel as evaluation data and the last one is a pool of these two. The results without any score normalization are marked as "baseline". For all experiments with adaptive score normalization, we used the top 200 files. By inspecting Figure 1, we can make several conclusions:

- S-norm produces the best results with 30% relative improvement, T-norm is behind and Z-norm is worse. ZT-norm produces the worst results in this setup.
- if matched data are present in the cohort, the results are much better, if they are not present, there is only a slight improvement over the baseline
- if matched data are present in the cohort then the adaptive score normalization is always better than using all data, because it selects the "correct" cohort
- adaptive S-norm2 used in [20] performs about the same as adaptive S-norm1. We also tried to apply adaptive Z-norm on the top of adaptive S-norm2 [20], but our results did not show any improvement.

Figure 2 shows $\mathrm{DCF}^{\mathrm{min}}$ as a function of the size of selected cohort in adaptive S-norm1. As before, there are three different cohort sets. All curves have nice flat minima between 200-500, we prefer 200 for practical reasons. The same trend was observed also on other conditions. It is also clear that using all data from the cohort yield worse results.

There are few tricks learned during these experiments to eliminate outliers from the cohort. The cohort set has an assumption to contain only one file per speaker, which might be hard to ensure in reality. When designing the cohort set on data without speaker labels, it is better to run unsupervised speaker clustering [31] and take only one file from each cluster. When selecting the cohort scores, it is also advantageous to eliminate/reject outlier scores by setting a "safe" interval from minus to plus 4-5 times standard deviation around the mean, and reject all cohort data outside.

Figure 3 compares adaptive S-norm1 with top 200 files in the cohort for different conditions and different cohorts. By examining all conditions and cohorts we conclude that:

- The cohort should contain the same languages as the evaluation set - channel match is not enough. This can be seen on the sre16evl,all condition with SRE16unlab (contain the

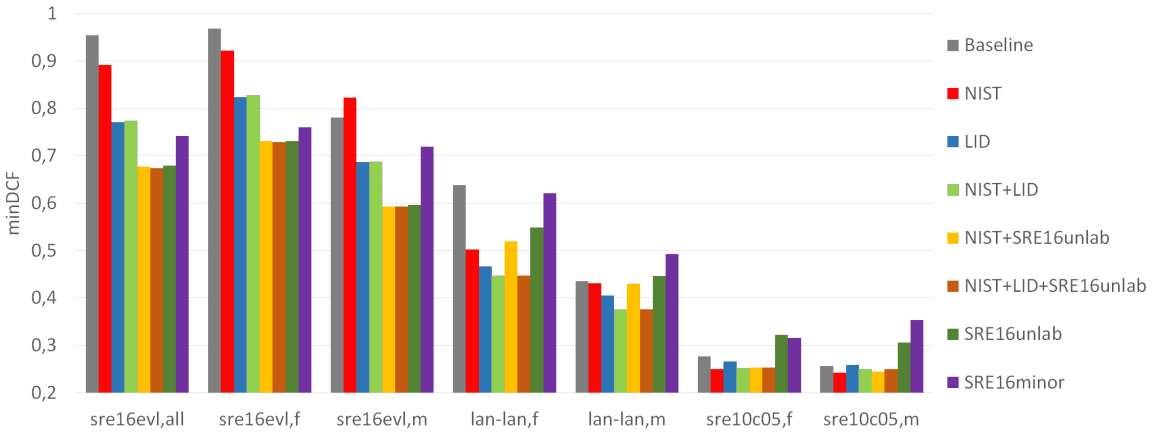

Figure 3: *Comparison of adaptive s-norm with different cohort sets with top 200 selected files.*

Table 2: *Results for all pooled NIST SRE 2016 trials for different cohorts, and adaptive s-norm with top 200 selected files.*

Cohort Set	$\mathrm{DCF}^{\mathrm{min}}$
Baseline (no norm)	0,9538
NIST	0,8922
LID	0,7712
NIST + LID	0,7739
NIST + SRE16unlabeled	0,6771
NIST + LID + SRE16unlabeled	**0,6733**
SRE16 unlabeled	0,6797
SRE16 minor	0,7418

same languages as the evaluation set) and SRE16minor (similar channel as the evaluation data, but different languages).

- if the cohort is too different, the performance can even degrade (see sre10c05 conditions and SRE16unlab and SRE16minor cohort sets).
- big set of LID data in the cohort is generally helpful especially in multilingual conditions.

Table 2 presents the details of $\mathrm{DCF}^{\mathrm{min}}$ for SRE 2016 all trials.

The following step was to analyze which files were selected to the cohort during the adaptive S-norm1 process. We run this analysis with top 50 files with NIST+sre16unlabeled cohort set. The first part of Table 3 shows the results of sre16evl,all and the languages with the largest coverage in the selected cohorts. SRE16unlab is present in the cohorts with 51% and 64% for male and female trials respectively which is obvious because it contains matched languages (Cantonese and Tagalog) and channel data. The second is Cantonese which is the target language (in this case from NIST data). English is the third most probably because it has a lot of files in the cohorts under various conditions. The following Mandarin, Vietnamese, Thai and Tagalog are languages quite close to target ones, with Tagalog being one of the target languages.

The sre16evl,all set contains 63.9% of Cantonese and 36.1% Tagalog data for male speakers. The selected cohort sets contain the data with same language from SRE16unlabeled[4] (51%), Cantonese (10.2%) and Tagalog (0.3%), the total evaluation language match is therefore 61.5%. For females, this number is 74.5%.

The second part of Table 3 describes PRISM lan-lan condition with true distribution of languages given in brackets. There is again a strong agreement between what language is selected

[4]Contain Cantonse and Tagalog data, but there are no labels for this data, we do not know precise numbers

Table 3: *Per-language analysis of files selected to the cohorts in adaptive score normalization. The numbers in brackets show real distribution of languages in the set in %.)*

Condition	Language	Male $\mathrm{DCF}^{\mathrm{min}}$	Female $\mathrm{DCF}^{\mathrm{min}}$
sre16evl,all	SRE16unlab	51.1	64.3
	Cantonese	10.2 (63.9)	9.0 (43.7)
	English	7.7	4.6
	Mandarin	5.8	2.9
	Vietnamese	2.4	4.8
	Arabic	5.2	2.9
	Thai	1.0	3.3
	Tagalog	0.3 (36.1)	1.2 (56.3)
lan-lan	English	61.7 (72.3)	55.4 (71.9)
	Mandarin	11.1 (17.2)	14.5 (25.1)
	Arabic	6.3 (3.6)	2.4 (2.8)
	Russian	1.6 (2.6)	5.5 (9.4)
	Thai	1.0 (4.4)	6.0 (13.0)
sre10c05	English	91.0 (100)	93.8 (100)
	other	$<$1.0	$<$1.0

to cohorts and the true percentage. The same is valid also for the last English condition sre10c05 where more than 90% of data selected to cohorts comes from English.

Globally, for all test data and all languages, we obtain in average 68% agreement between the language of enrollment and test files and language selected to the cohorts. There is also strong agreement in gender – the files in selected cohorts match in 92% cases the gender of the evaluation condition.

5. Conclusions

Our paper shows the outcomes of analysis of score normalization techniques. The results are obtained mainly on NIST SRE 2016, but we also report results on NIST SRE 2010 and multilanguage condition from PRISM set. The analysis shows that using adaptive symmetric score normalization (s-norm) performs the best with 30% relative improvement. The best results were achieved by selecting 200-500 top scoring files to create a speaker-dependent cohort. Further analysis shows that the selected cohorts match in 68% the language and in 92% the gender of the enrollment and test recordings. Next, our experimental results suggest that the general score normalization cohort should be a pool of several languages and channels and if possible, its subset should contain the data from the target domain (language and channel).

6. References

[1] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification." in *Eurospeech*, 1997.

[2] D. Reynolds, M. Zissman, T. Quateri, G. O. Leary, and B. Carlson, "The effects of telephone transmission degradations on speaker recognition performance." in *ICASSP*, May 1995.

[3] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus." in *ICASSP*, May 1996.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19 – 41, 2000.

[5] P. Kenny, N. Dehak, R. Dehak, V. Gupta, and P. Dumouchel, "The role of speaker factors in the NIST extended data task," in *Proceedings of the Speaker and Language Recognition Workshop (IEEE-Odyssey 2008)*, Stellenbosch, South Africa, Jan. 2008.

[6] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, July 2008.

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

[8] H. Aronowitz, D. Irony, and D. Burshtein, "Modeling intra-speaker variability for speaker recognition," in *in Proc. Interspeech*, 2005.

[9] M. Ben, R. Blouet, and F. Bimbot, "A Monte-Carlo method for score normalization in automatic speaker verification using kullback-leibler distances," in *ICASSP*, 2002, pp. 689–692.

[10] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Speaker verification using speaker- and test-dependent fast score normalization," *Pattern Recognition Letters*, vol. 28, pp. 90 – 98, 2007.

[11] P. Kenny, "Bayesian speaker verification with heavy–tailed priors," keynote presentation, Proc. of Odyssey 2010, Brno, Czech Republic, June 2010.

[12] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification," in *Odyssey The Speaker and Language Recognition*, Brno, Czech Republic, 2010.

[13] V. R. Apsingekar and P. L. D. Leon, "Speaker verification score normalization using speaker model clusters," *Speech Communications*, vol. 53(1), pp. 110–118, January 2011.

[14] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalization methods in open-set speaker identification," in *Odyssey The Speaker and Language Recognition*, Toledo, Spain, 2004.

[15] Y. Zigel and A. Cohen, "On cohort selection for speaker verification," in *Eurospeech*, 2003.

[16] H. Aronowitz and V. Aronowitz, "Efficient score normalization for speaker recognition," in *ICASSP*, 2010.

[17] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *ICASSP*, 2005, pp. 741–744.

[18] Y. Zigel and M. Wasserblat, "How to deal with multiple-targets in speaker identification systems?" in *Proceedings of the Speaker and Language Recognition Workshop (IEEE-Odyssey 2006)*, San Juan, Puerto Rico, June 2006.

[19] Z. N. Karam, W. M. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *ICASSP*, 2011.

[20] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *INTERSPEECH 2011*, Florence, Italy, August 2011.

[21] O. Plchot, "Analysis and Description of ABC Submission to NIST SRE 2016," in *submitted to Interspeech 2017*, Stockholm, Sweden, 2017.

[22] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Eurospeech*, September 2005.

[23] D. Colibro, C. Vair, E. Dalmasso, K. Farrell, G. Karvitsky, S. Cumani, and P. Laface, "Nuance - Politecnico di Torinos 2016 NIST Speaker Recognition Evaluation System," in *ICSLP*, Stockholm, Sweden, August 2017.

[24] "The 2016 NIST speaker recognition evaluation plan (sre16)," https://www.nist.gov/file/325336.

[25] "The 2010 NIST speaker recognition evaluation plan (sre10)," https://www.nist.gov/document-11909.

[26] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," *https://code.google.com/p/prism-set/*, 2012.

[27] Z. Jančík, O. Plchot, N. Brummer, L. Burget, O. Glembek, V. Hubeika, M. Karafiát, P. Matějka, T. Mikolov, A. Strasheim, and J. Černocký, "Data selection and calibration issues in automatic language recognition - investigation with BUT-AGNITIO NIST LRE 2009 system," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010, pp. 215–221.

[28] O. Plchot, P. Matějka, R. Fér, O. Glembek, O. Novotný, J. Pešán, K. Veselý, L. Ondel, M. Karafiát, F. Grézl, S. Kesiraju, L. Burget, N. Brummer, P. du Albert Swart, S. Cumani, H. S. Mallidi, and R. Li, "BAT System Description for NIST LRE 2015," in *Proceedings of Odyssey 2016, The Speaker and Language Recognition Workshop*, vol. 2016, no. 06, 2016, pp. 166–173.

[29] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," vol. PP, no. 99, pp. 1 –1, 2010.

[30] P. Matějka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocký, "Full-covariance UBM and heavy-tailed PLDA in I-vector speaker verification," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. IEEE Signal Processing Society, 2011, pp. 4828–4831.

[31] S. Shum, D. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *The Speaker and Language Recognition Workshop: Odyssey*, Joensuu, Finlan, June 2014.