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

# Improving Robustness of Speaker Recognition to New Conditions Using Unlabeled Data

*Diego Castan*[1], *Mitchell McLaren*[1], *Luciana Ferrer*[2], *Aaron Lawson*[1], *Alicia Lozano-Diez* [3]

[1]Speech Technology and Research Laboratory, SRI International, California, USA
[2]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[3]Audias-ATVS, Universidad Autonoma de Madrid, Madrid, Spain

{dcastan,mitch,aaron,alozano}@speech.sri.com, lferrer@dc.uba.ar

## Abstract

Unsupervised techniques for the adaptation of speaker recognition are important due to the problem of condition mismatch that is prevalent when applying speaker recognition technology to new conditions and the general scarcity of labeled 'in-domain' data. In the recent NIST 2016 Speaker Recognition Evaluation (SRE), symmetric score normalization (S-norm) and calibration using unlabeled in-domain data were shown to be beneficial. Because calibration requires speaker labels for training, speaker-clustering techniques were used to generate pseudo-speakers for learning calibration parameters in those cases where only unlabeled in-domain data was available. These methods performed well in the SRE16. It is unclear, however, whether those techniques generalize well to other data sources. In this work, we benchmark these approaches on several distinctly different databases, after we describe our SRI-CON-UAM team system submission for the NIST 2016 SRE. Our analysis shows that while the benefit of S-norm is also observed across other datasets, applying speaker-clustered calibration provides considerably greater benefit to the system in the context of new acoustic conditions.

**Index Terms**: Score Normalization, Score Calibration, Trial-based Calibration, NIST SRE16

## 1. Introduction

Speaker recognition systems compare voice samples (a.k.a., a trial) to provide a measure of voice similarity. This measure, or trial score, from the system may be an arbitrary number that provides information only when measured relative to other scores. A more appropriate output of a speaker recognition system is a calibrated log-likelihood ratio (LLR) [1]. The LLR provides a self-contained ratio of the probability of the voices being from the same speaker versus the alternate hypothesis of them being from different speakers. That is, the LLR is meaningful on its own (i.e., a LLR of 2 means the same-speaker hypothesis is 100 times more likely than the alternate hypothesis). To obtain well-calibrated LLRs from a system, the system training data, including calibration training data, must closely represent the conditions of the trial being evaluated [2]. In practice, this prerequisite is often unfulfilled due to acoustic or other differences between system-development data and the audio observed during use; a problem known as condition mismatch. Several approaches have been developed to reduce condition mismatch, and these can be categorized as *unsupervised* or *supervised* methods.

Unsupervised methods do not require ground-truth speaker labels for training or application. These are mostly concerned with an array of score-normalization approaches [3, 4, 5] that leverage an "impostor" cohort (a held-out data set) to generate a speaker model- or test-centric impostor distribution from which statistics are derived and used to normalize a trial score to fit within a zero-mean, unit-variance score distribution. These techniques aim to normalize score distributions across varying conditions to ease the load on subsequent processes, such as threshold application, speaker ranking, or calibration. In contrast, supervised methods for speaker recognition require speaker labels during training in order to observe and model both different- and same-speaker trial scores. Techniques that fall under this umbrella include domain adaptation of the backend classifier to directly improve speaker discrimination [6, 7, 8], score calibration to enable better system practical application [1], and dynamic calibration methods that utilize test-time information to make better calibration choices [2]. In the absence of speaker labels, speaker clustering can be used to create pseudo-speaker labels, thus enabling the application of supervised techniques without having hand-annotated data [6].

Score normalization does not calibrate scores. As such, even after normalization, calibration is an essential process to generate actionable LLRs on which decisions can be made. Until the recent NIST SRE16, studies on calibration focused on leveraging ground-truth speaker labels, while domain adaptation had already been considered in the "unsupervised" case using clustered speaker labels [6, 7, 8]. Symmetric score normalization (S-Norm) has been found to be useful in the context of cosine-based scoring [4, 5], and several submissions to the SRE16 were shown to benefit from this approach in the context of probabilistic linear discriminant analysis (PLDA)-based scoring when coping with the SRE16 data-mismatch problem where in-domain impostor data could be leveraged. Considering these findings, we address here the question of whether S-Norm also provides additional system robustness to mismatch when applied to non-SRE databases (assuming unlabeled data is available), and whether this data can also be leveraged for calibration.

Usually, it is easy to find unlabeled data matching the conditions presented on the test. It is common to cluster that unlabeled data into 'pseudo-speakers' and use those labels into two main techniques: domain adaptation and calibration. In the NIST SRE16, unlabeled data was provided that represented the conditions of the evaluation data [9] and to constrain the scope of this paper, we focus our research on the calibration and nor-

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

http://dx.doi.org/10.21437/Interspeech.2017-605

malization techniques, leaving the domain adaption techniques for further studies. Therefore, this paper studies and compares those techniques used in the SRE16 such as, S-Norm, adaptive S-Norm (AS-Norm) [6], linear logistic regression, or trial-based calibration, which were deemed helpful for dealing with the mismatch between development and evaluation data. We compare these techniques on several distinctly different databases to determine which techniques, or combination thereof, generalize to these new environments when unlabeled data also exists for system adaptation.

# 2. Normalization

Score normalization consists of a set of methods for improving robustness against variability by shifting and scaling scores to fit a predefined impostor score distribution. The aim is to reduce the variability between trials, while subsequently enables more reliable decisions through the application of a decision threshold or ranking of trial scores. This is accomplished by normalizing the raw trial score (typically LLRs from PLDA) relative to a set of cohort models with respect to the speaker model and/or test sample, thereby transforming the scores to remove the speaker- or test-dependence on the expected impostor score distribution. Normalization fits impostor scores to a normal zero-mean, unit standard-deviation Gaussian distribution by using $S' = \frac{S-\mu_i}{\sigma_i}$ where $\mu_i$ and $\sigma_i$ are estimations of the mean and the standard deviation of the impostor scores generated from the impostor cohort in one of several ways. Each normalization technique defines a different method for estimating these statistics. These techniques include Z-Norm for speaker-centric normalization, T-Norm for test-centric normalization, or a combination of both for ZT-Norm [3]. Speaker- and test-adaptive versions of these were also proposed to better normalize for trial-specific differences [10].

In the i-vector paradigm, the trials are symmetric (i.e., the same LLR results irrespective of two samples being assigned as enrollment and test samples or vice versa). Accordingly, symmetric score normalization (S- Norm) was developed in the context of cosine scoring of i-vectors [5, 4]. The NIST SRE16 found that both S-Norm and adaptive S-Norm (AS-Norm) improved robustness when condition mismatch existed between system PLDA backend training data and evaluation data. This finding motivated us to look deeper into how robustly the S-Norm techniques enhance robustness to unseen conditions.

## 2.1. Symmetric normalization (S-Norm)

S-Norm uses a single cohort that contains speech segments from speakers not expected to be observed in the evaluation data. Its respective S-Norm parameters are the mean and standard deviation of the scores between the given utterance (test or enrollment) and the entire cohort. The normalized score is given by,

$$S' = \frac{S - \mu_t}{\sigma_t} + \frac{S - \mu_m}{\sigma_m}, \tag{1}$$

where $\mu_t$ and $\sigma_t$ are estimated from the score distribution of the test segment against the cohort, and $\mu_m$ and $\sigma_m$ as estimated from trial score distribution of the model (i.e., enrollment) segment against the cohort. S-Norm is expected to be most effective when the conditions of the impostor cohort are representative of the evaluation trial conditions.

## 2.2. Adaptive symmetric normalization

We also used a dynamic or adaptive variant of S-Norm, referred to as Adaptive S-Norm (AS-Norm), which adjusts to the conditions of the trial at hand [10]. This approach selects a subset of samples from the held-out cohort to compute the normalization parameters. The selection is done based on the similarity between the trial sides and the cohort samples. It creates a vector of scores pairing up the held-out samples against a set of other held-out samples. It then does the same for the sides in the trial and selects the samples for which the vector of scores is closest to the one for each of the sides. In other words, the criteria defines similarity between two samples based on the scores that the system produces when testing these samples against impostor samples. The samples that are selected for normalization could then be either from similar conditions, similar speakers, or both, to the samples in the trial.

# 3. Calibration

Whether scores are normalized or not, system calibration remains an important process. While normalization modifies the distribution of the scores to provide more robustness against different conditions, it does not produce calibrated LLRs. A speaker recognition system is considered to have most utility when calibrated log-likelihood ratios (LLR) are its output [11]. In many circumstances, even the most speaker-discriminative system will be of limited application unless it is well calibrated. The role of calibration is transforming scores (normalized or otherwise) into interpretable LLRs [11]. By interpretable, we refer to the value of a calibrated LLR, which relays the weight of evidence with respect to the same-speaker and different-speaker hypotheses.

Once the scores are properly calibrated, the optimal threshold for a certain cost function can be determined using Bayes decision theory. Scores are usually transformed by a monotonically increasing function with parameters that must be trained. If the function consists of few parameters, then the risk of over-fitting is low [12]. Linear logistic regression is generally used for calibrating speaker recognition systems due to its ability to generalize well and the availability of toolkits such as FoCal [13] and BOSARIS [14]. Alternate methods of calibrating systems have also been proposed that consider meta-data [15, 16]. More recently, Trial-based Calibration (TBC) was proposed in [2] as a dynamic calibration technique aimed at providing robustness to unseen and widely varying conditions by leveraging external information from the enrollment and test samples to select appropriate calibration training data for the trial.

In contrast to normalization, the calibration process requires knowledge of both the expected target and impostor trial distributions, typically generated using a representative dataset with speaker labels.

## 3.1. Logistic regression

Perhaps the most commonly used calibration technique in the field of speaker recognition is logistic regression [13]. A calibration model (shift and bias) is learned by using linear logistic regression from a large pool of trials deemed to be representative of the end-use conditions. A single model is typically trained using all development data and requires speaker labels to model target versus impostor score distributions.

## 3.2. Trial-Based calibration (TBC)

Trial-based calibration [2] was recently developed to cope with dramatic mismatch between system calibration data, enrollment speech, and test speech. This dynamic calibration approach uses information about the conditions in both sides of the trial to adaptively calibrate the trial. This is accomplished using universal audio characterization (UAC) [17], which extracts a metadata vector (UAC vector) for each trial side. This vector indicates the likelihood of predefined classes such as gender, duration class, and bitrate (as used in this work based on SRE16 development set results). These trial UAC vectors are compared to the UAC vectors extracted from a candidate calibration dataset to downselect the most representative trials from which to learn the trial-dependent calibration model. Here, we downselected to 500 target trials as tuned on the SRE16 development set. Calibration parameters are learned through simple logistic regression using this small, highly relevant calibration dataset with which the trial score is calibrated. In this way, the conditions in the calibration model for the trial are more representative than would be the case for a global calibration model learned on all data and result in better calibration across mixed conditions [2]. For more details on TBC, readers are referred to [2].

In the current context, we extended this technique in two ways: (1) using clustered speaker labels instead of ground truth and (2) using in-domain data as opposed to mismatched data. Being a dynamic process, TBC is much like AS-Norm in that it considers information from both sides of the trial. However, rather than normalizing the impostor score distribution using downselected segments chosen by speaker similarity, TBC calibrates to proper LLRs and uses information outside the task of speaker comparison to make the choice of calibration segments.

### 3.3. Speaker clustering

To leverage supervised methods in the context of unlabeled data, pseudo-speaker labels must first be estimated. For this purpose, we employed speaker clustering.

We clustered the unlabeled data into pseudo-speakers using the following procedure for each evaluated dataset. First, the data was divided into language groups (minor/major for SRE16, and five different languages for RATS). The clustering procedure was done separately for each of these language sets, and i-vectors were created for each sample in each set. A PLDA model trained on the training data from the closed condition was used to generate all-vs.-all scores for the i-vectors in each set. These scores were then transformed into a distance matrix. The distances were computed as the opposite of the log-likelihood ratios (LLR) obtained with PLDA, shifted by the maximum LLR obtained for any pair of samples. This way, the minimum distance was 0. Finally, hierarchical clustering with an average linkage method was used to generate a clustering tree. The tree was then pruned by ensuring that each cluster had a cophenetic distance no greater than a certain threshold, $t$. The value of $t$ that optimized clustering performance (as measured by the adjusted rand index) on the SRE16 development speakers was used to cluster the unlabeled segments. This resulted in 615 clusters (532 from the major set and 83 from the minor set) of sizes ranging from 1 to 187 for SRE16.

## 4. System Description

The speaker recognition system used in this study was based on our primary submission for the SRE16 evaluation. It involved a four-way score-level fusion of a single DNN-based system leveraging the hybrid alignment framework [18] and three traditional universal background model (UBM) i-vector systems based on mel-frequency cepstral coefficients (MFCC), power-normalized cepstral coefficients (PNCC) [18], and pitch-based PROS features [19]. Clustering was performed using the PLDA scores obtained from employing an i-vector-fusion of three of the subsystems (all except for MFCC) for improved discrimination during clustering.

The DNN training data was Fisher and Switchboard 1 data, per the restrictions of the fixed training condition of SRE16. The UBM and i-vector subspaces were trained using all available clean telephone data in the Mixer datasets only. Other datasets and microphone conditions were found to degrade performance on the development set. The PLDA models in all subsystems were trained using Mixer telephone data, and Mixer microphone data degraded with reverb or compression. For a more complete description of the system, including our exploration of PLDA domain adaptation, we refer the reader to [21].

## 5. Experiment Protocol

All the experiments in this section were performed using the “fixed” training condition in SRE16. This condition limits system training to the data provided from the 2016 corpus collection, from previous SREs, and to the Switchboard corpora that contain transcripts. No other auxiliary data was allowed. The data were composed of telephone conversations spoken in Tagalog and Cantonese (referred to as the major languages) and Cebuano and Mandarin (referred to as the minor languages). The development set contains data from the two minor languages, while the evaluation data is based on the two major languages. This paper presents results only for the evaluation dataset since the results on the development set were very noisy and present large confidence intervals due to the limited amount of data and our focus is on the generalization of techniques to alternate datasets. An unlabeled set of approximately 2200 conversation sides was provided for SRE16, matching the evaluation conditions and containing labels only regarding which language group the audio was from.

We also benchmarked on part of the DARPA RATS speaker recognition corpus and the Speakers in the Wild (SITW) corpus. For RATS corpus, we used the original (source) telephone segments in the collection of push-to-talk transmissions for the DARPA RATS speaker recognition dataset [22]. The telephone data was sourced from five non-English languages (Arabic levantine, Farsi, Dari, Pashto and Urdu) and contained speech from 336 speakers. As in SRE16, no cross-language trials were considered. We truncated each of these files to include 10 to 60 seconds of speech, then divided the dataset by speaker label into “unlabeled” and “eval” splits, each having equal language distribution. These splits had 842/87 and 9624/249 segments/speakers, respectively. For the eval set we created six cuts per original segment with durations between 10 and 60 seconds of speech. For the unlabeled set, a random duration was selected for each original segment.

The SITW dataset is composed of nearly 300 individuals across clean interview, red carpet interviews, stadium conditions, outdoor conditions, and multi-speaker scenarios. Each individual also has speech acquired from unedited camcorder or cellphone footage. All noise, reverberation, compression and other artifacts in the corpus are natural characteristics of the original audio. We refer to [23] for more information about the data collection and a detailed description of the different conditions. For the purpose of this study, we constrained the data to single-speaker enrollment and tests (i.e., the core-core condition) and used the development partition as “unlabeled” data, and report results on the “eval” set.

# 6. Results

This section compares normalization and calibration techniques in the context of SRE16 when leveraging unlabeled in-domain data. We also benchmark these approaches on other databases to determine whether trends generalize to other data. We benchmark the overall performance with the well-known Cllr and EER metrics, and the calibration performance with minDCF and ActDCF.

## 6.1. Results using true speaker labels for calibration

We start our experiments by focusing on the unsupervised techniques of S-Norm and AS-Norm when we calibrate with true speaker labels. These approaches were benchmarked on SRE16, RATS and SITW corpora with results given in Table 1. For this analysis, we calibrated scores with true speaker labels from the SRE16 development data and the ground-truth speaker labels of the RATS and SITW "unlabeled" data. SRE16 showed a 17% and 19% relative gain in Cllr, and actDCF with equal cost respectively, through application of either S-Norm or AS-Norm over the raw scores with global calibration. Although TBC with S-Norm provides the best EER and minimum costs for SRE16, there is greater miscalibration (difference between actual and minimum costs) when using TBC compared to global calibration. These results demonstrate the substantial contribution that S-Norm makes in the context of SRE16.

On the RATS data, S-Norm gives a significant improvement over unnormalized scores for all metrics except Cllr. As in SRE data, AS-Norm does not give significant improvements over simple S-Norm. For this data, TBC gives a significant improvement for unnormalized scores and a somewhat less consistent improvement for S-Norm scores. Finally, for SITW data, S-Norm with TBC is consistently better for all the metrics

Overall, these results suggest that S-Norm is a viable option to improve the robustness of a system across datasets, assuming in-domain data is available for the purpose, and additionally applying TBC with labeled in-domain data provides better discriminative power, but this is not always correlated with calibration performance.

Table 1: *Application of S-Norm and AS-Norm using unlabeled data on SRE16, RATS and SITW databases. Calibration was performed using the labeled development set.*

| | | | | | Equal Cost | |
|---|---|---|---|---|---|---|
| | CAL | NORM | CLLR | EER | MDCF | ADCF |
| SRE16-Eval | global | Raw | 0.508 | 12.98 | 0.259 | 0.308 |
| | global | S-Norm | 0.419 | 12.25 | 0.241 | 0.247 |
| | global | AS-Norm | **0.412** | 12.10 | 0.238 | **0.241** |
| | TBC | Raw | 0.466 | 12.60 | 0.250 | 0.280 |
| | TBC | S-Norm | 0.477 | **11.55** | **0.225** | 0.299 |
| RATS | global | Raw | 0.228 | 4.74 | 0.093 | 0.126 |
| | global | S-Norm | 0.224 | 4.13 | 0.081 | **0.081** |
| | global | AS-Norm | 0.213 | 4.14 | 0.081 | 0.120 |
| | TBC | Raw | 0.190 | 4.35 | 0.086 | 0.094 |
| | TBC | S-Norm | **0.173** | **3.97** | **0.078** | 0.087 |
| SITW | global | Raw | 0.273 | 7.83 | 0.155 | 0.157 |
| | global | S-Norm | 0.282 | 7.81 | 0.154 | 0.157 |
| | global | AS-Norm | 0.271 | 7.79 | 0.154 | 0.156 |
| | TBC | Raw | 0.251 | 7.31 | 0.144 | 0.144 |
| | TBC | S-Norm | **0.251** | **7.19** | **0.142** | **0.144** |

## 6.2. Results using pseudo-speaker labels for calibration

In this section we use pseudo-speaker labels obtained from clustering instead of true labels for calibration, and we use the "unlabeled" data for SRE16 calibration instead of the development set. In the previous section, we found no significant advantage from using AS-Norm instead of S-Norm. Therefore, we chose to optionally apply S-Norm in the following experiments as opposed to AS-Norm, due to the lower computational cost. To apply both S-Norm and calibration using the same dataset, S-Norm was first applied in a cross-validation manner before pooling the normalized scores on which calibration was then trained.

Table 2: *Completely unsupervised techniques global calibration and TBC with clustered speaker labels on SRE16, RATS and SITW databases.*

| | | | | | Equal Cost | |
|---|---|---|---|---|---|---|
| | CAL | NORM | CLLR | EER | MDCF | ADCF |
| SRE16-Ev. | global | Raw | 0.637 | 12.98 | 0.259 | 0.398 |
| | global | S-Norm | 0.523 | **12.25** | **0.241** | **0.258** |
| | TBC | Raw | **0.485** | 12.44 | 0.247 | 0.303 |
| | TBC | S-Norm | 0.691 | 18.33 | 0.356 | 0.405 |
| RATS | global | Raw | 0.195 | 4.74 | 0.093 | 0.100 |
| | global | S-Norm | 0.258 | 4.13 | 0.081 | 0.121 |
| | TBC | Raw | **0.189** | 4.58 | 0.089 | **0.092** |
| | TBC | S-Norm | 0.323 | **4.12** | **0.080** | 0.152 |
| SITW | global | Raw | 0.767 | 7.83 | 0.155 | 0.290 |
| | global | S-Norm | **0.484** | 7.81 | **0.154** | **0.211** |
| | TBC | Raw | 0.817 | 7.83 | 0.155 | 0.306 |
| | TBC | S-Norm | 0.566 | **7.80** | 0.155 | 0.224 |

Several different comparisons can be made from the results in Table 2. First, we consider performance of calibration in the absence of S-Norm. The application of clustered calibration tends to provide reasonable actDCF performance on SRE16 and RATS datasets, indicating suitability of the approach, however the calibration performance on SITW was considerable worse than use of groundtruth labels in Table 1. This may be attributed to worse clustering performance on the considerably varying conditions of SITW. In contrast to groundtruth labels, TBC with clustered labels did not benefit from the application of S-Norm with the exception of RATS EER and MDCF metrics, and EER in SITW.

While trends are not consistent across datasets, it would appear that either the application of S-norm and global calibration from clustered speakers, or the application of TBC and no normalization are the most feasible options when using unlabeled in-domain data.

# 7. Conclusions

This paper compared using unlabeled in-domain data during the normalization and calibration stages of a speaker recognition system, across several different datasets. The unsupervised approaches of symmetric normalization (S-Norm) and adaptive S-Norm were found to offer robustness in some instances, while never reducing system performance when supervised calibration was applied. The application of global calibration using clustering to generate pseudo-speaker labels was found to provide reasonable calibration performance when clustering performance was sufficient for the conditions. Introducing S-Norm as a pre-processing step to global calibration with the same unlabeled data showed the best minDCF performance, but at a much greater cost to calibration performance. In the absence of labeled in-domain data, the application of S-Norm followed by global calibration demonstrated the most robustness across datasets, with S-Norm improving overall speaker discrimination. Future work will investigate using disjoint datasets for the normalization and calibration tasks and better generalization of TBC when in-domain data is available.

# 8. References

[1] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.

[2] M. McLaren, A. Lawson, L. Ferrer, N. Scheffer, and Y. Lei, "Trial-based calibration for speaker recognition in unseen conditions," in *Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.

[3] 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.

[4] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification." in *Odyssey*, 2010, p. 16.

[5] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques." in *Odyssey*, 2010, p. 15.

[6] S. Shum, D. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Proc. Odyssey*, 2014.

[7] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 2014, pp. 378–383.

[8] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4032–4036.

[9] *The NIST Year 2016 Speaker Recognition Evaluation Plan*, 2016, available at www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016.

[10] D. Sturim and D. Reynolds, "Speaker adaptive cohort selection for t-norm in text-independent speaker verification," in *Proc. ICASSP*, 2005.

[11] N. Brummer and D. Van Leeuwen, "On calibration of language recognition scores," in *Odyssey: Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[12] J. Villalba, "Advances on speaker recognition in non-collaborative environments," Ph.D. dissertation, University of Zaragoza, 2014.

[13] N. Brummer, *Focal toolkit*, 2006, available at sites.google.com/site/nikobrummer/focal.

[14] ——, *Bosaris toolkit*, 2010, available at sites.google.com/site/bosaristoolkit/.

[15] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification," in *Proc. ICASSP*, 2008.

[16] V. Hautamaki, K. Lee, T. Kinnunen, B. Ma, and H. Li, "Regularized logistic regression fusion for speaker verification," in *Proc. Interpseech*, 2011.

[17] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, "A unified approach for audio characterization and its application to speaker recognition." in *Odyssey*, 2012, pp. 317–323.

[18] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4101–4104.

[19] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," in *Proc. Interpseech*, 2013.

[20] M. McLaren and D. Van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 755–766, 2012.

[21] M. McLaren, L. Ferrer, D. Castan, A. Lawson, and A. Lozano-Diez, "The SRI-CON-UAM NIST 2016 SRE system description," in *Proc. SRE16 Workshop*, 2016.

[22] K. Walker and S. Strassel, "The RATS radio traffic collection system." in *Odyssey*, 2012, pp. 291–297.

[23] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The Speakers In The Wild (SITW) speaker recognition database," in *Proc. Interspeech*, 2016.