INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

On the Issue of Calibration in DNN-based Speaker Recognition Systems

Mitchell McLaren[1], *Diego Castan*[1], *Luciana Ferrer*[2], *Aaron Lawson*[1]

[1]Speech Technology and Research Laboratory, SRI International, California, USA
[2]Departamento de Computación, FCEN, Universidad de Buenos Aires and CONICET, Argentina
{mitch,dcastan,aaron}@speech.sri.com, lferrer@dc.uba.ar

Abstract

This article is concerned with the issue of calibration in the context of Deep Neural Network (DNN) based approaches to speaker recognition. DNNs have provided a new standard in technology when used in place of the traditional universal background model (UBM) for feature alignment, or to augment traditional features with those extracted from a bottleneck layer of the DNN. These techniques provide extremely good performance for constrained trial conditions that are well matched to development conditions. However, when applied to unseen conditions or a wide variety of conditions, some DNN-based techniques offer poor calibration performance. Through analysis on both PRISM and the recently released Speakers in the Wild (SITW) corpora, we illustrate that bottleneck features hinder calibration if used in the calculation of first-order Baum Welch statistics during i-vector extraction. We propose a hybrid alignment framework, which stems from our previous work in DNN senone alignment, that uses the bottleneck features only for the alignment of features during statistics calculation. This framework not only addresses the issue of calibration, but provides a more computationally efficient system based on bottleneck features with improved discriminative power.
Index Terms: speaker recognition, mismatch, calibration, deep neural network, bottleneck features

1. Introduction

In recent years, deep neural networks (DNN) have been widely applied to speech applications, including speaker recognition [1, 2, 3]. The first applications of "senone" DNNs to this task that provided a large improvement over alternate technology were in [1, 4]. DNNs were applied from the field of automatic speech recognition to provide class posteriors for i-vector extraction based on tied tri-phone states, or "senones," instead of component posteriors from a Universal Background Model (UBM) that was trained to cluster acoustically similar sounds in an unsupervised manner. This approach allowed direct comparison of the way in which two different speakers pronounce the same phone, and resulted in a major advancement in speaker recognition under telephony conditions. The same trend, however, was not observed on microphone or alternate channel conditions [5], although methods to improve robustness to non-telephone channels have been proposed [6]. More recently, bottleneck (BN) features extracted from DNNs trained to predict senones were shown to be very successful in the related field of language recognition [7, 8, 9]. These BN features were then applied to speaker recognition as a single feature and in combination with MFCCs [6]. This latter BN+MFCC combination was found to be one of the most robust DNN-based options for speaker recognition when evaluated across a number of conditions [6]. Studies to date have focused on the ability of DNN-based speaker recognition to provide state-of-the-art improvements in technology by observing their discrimination power in relatively homogeneous conditions. However, the aspect of calibration and performance under varying trial conditions has, to the best of our knowledge, not yet been investigated.

Calibration is a key aspect of any system and is typically applied as a transformation to system scores with the aim of producing proper (calibrated) log-likelihood ratios (LLR) [10]. A calibrated LLR represents the "strength of evidence" for the hypothesis of a same-speaker trial vs. a different speaker trial. Before applying calibration, parameters of a calibration model must be learned using a development set of same and different-speaker trial scores. One major difficulty with calibration is that the development set should consist of scores that are representative of the end use case; this requirement is not always attainable due to data scarcity. Further, when more than one condition is encountered by the system (such as telephone *or* microphone test samples), a single calibration model learned from the pooled-condition trial scores is not always a suitable solution [11]. For a verification system where a decision threshold will be applied, properly calibrated scores will allow application of a *single* threshold which should be optimal across all the trials. However, with mixed conditions, this calibration is difficult to obtain. In order to achieve this, a system should produce score distributions that are not sensitive to changes in channel or acoustic conditions.

In this work, we provide a study of calibration performance for the current DNN-based systems against the MFCC i-vector system. We demonstrate that of the DNN-based systems, the BN+MFCC system exhibits considerable condition dependence in score distributions. The use of BN features in the calculation of the first-order statistics was identified as the cause of this dependence, a finding which led to our proposal of a new hybrid alignment framework. One instance of this framework uses the BN features exclusively to produce alignments for the speaker identification (SID) features in the calculation of first-order statistics or accumulators for i-vector extraction, much like our DNN/i-vector approach [1]. When evaluated on both PRISM and the recently released Speakers in the Wild (SITW) databases, this framework is shown to provide considerably better calibration and discrimination performance as compared to the BN+MFCC system architecture.

2. Background

This section provides the technical background of the DNN approaches considered in this work. Additionally, we give an overview of linear calibration for speaker recognition.

http://dx.doi.org/10.21437/Interspeech.2016-1134

2.1. DNNs in Speaker Recognition

The successful application of senone DNNs to speaker recognition has focused on the i-vector framework [1, 2, 3]. Specifically, the DNN has been used in one of two ways: to extract bottleneck features (BN) as input into the traditional UBM-based i-vector framework [7, 11], or in a DNN/i-vector approach in which the posteriors of the output layer of the DNN are used to align an alternate set of features — hereafter referred to as SID features — in the calculation of the first-order Baum-Welch statistics [1].

The traditional i-vector framework, as it was proposed in [12], used a single feature to extract both the alignments and first order statistics. Specifically, a UBM was first trained to cluster MFCCs into unsupervised clusters. Then, posteriors for SID feature alignment were generated by simply decoding the feature through the UBM. This architecture provided major advancement on prior technology, namely joint factor analysis [13], and has since been widely adopted in the community.

Recently, the role of the UBM was replaced in the DNN/i-vector paradigm to provide major improvements, particularly in telephone-based speech [1]. In this paradigm, the DNN was trained to predict senones (commonly used for automatic speech recognition). The senone posteriors (with silence senones removed) were then used to replace UBM component posteriors when aligning SID features during statistics calculation. This provided a means of directly comparing the way two speakers pronounce the same phone from the same relative point (a supervised senone) of the analogous UBM, instead of from a cluster learned through unsupervised clustering of acoustic sounds.

More recent was the application of BN features extracted from a DNN to the task of SID [14]. This application was motivated by the significant benefit derived from these features in the field of language recognition [7]. Bottleneck features can be considered a phonetic representation from the DNN that is extracted as a set of linear activations from a hidden layer with a relatively small number of nodes compared to alternate layers (80 compared to 1200 in this work). These features, when concatenated with traditional MFCCs, can be used in the traditional i-vector paradigm. These BN+MFCC features have been shown to provide performance that is competitive with the DNN/i-vector approach, with the advantage of offering flexibility in the UBM/i-vector design. This flexibility can, subsequently, translate to a computational advantage. One disadvantage is the large feature dimension (80D + 60D) and the phonetic-dependence of the i-vectors [15], since the content of the BN features are used in the calculation of the first-order statistics. In this work, we focus on the differences between these approaches in terms of calibration performance as compared with the traditional MFCC i-vector framework.

2.2. Calibration of Speaker Recognition Systems

Transforming scores to likelihood ratios via calibration removes the arbitrary nature of system scores and allows information for a trial to be contained in a single number [10, 11]. The LLR indicates the support for a same-speaker hypothesis versus a different-speaker hypothesis. Linear logistic calibration is a simple transformation of a score, s, as $s_{cal} = s\alpha + \beta$. The calibration parameters α and β are typically learned from a development calibration set of trials which are representative of the conditions for which the system is intended to operate by maximizing the likelihood under the assumption that posterior probabilities are given by a logistic function of the scores. When

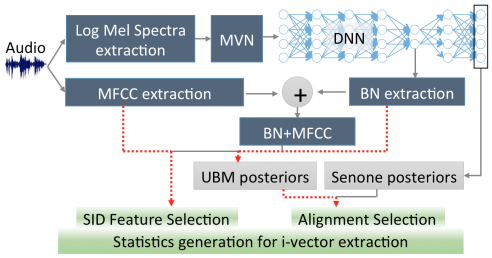

Figure 1: Extraction of features from audio, including bottleneck features and their concatenation with MFCCs and how they can be used as SID features, and to calculate posteriors via a UBM for the purpose of statistics generation. The alternative DNN/i-vector framework relies on senone posteriors for this alignment process.

accurate representation is not guaranteed, either through subtle changes in acoustic environment or considerable mismatch in terms of channel, system calibration is similarly not guaranteed. In order to use this simple calibration model in a manner that can withstand any degree of mismatch, the score distributions output from the system would ideally exhibit limited sensitivity to the trial conditions. As we will show later in this work, this sensitivity can have quite a dramatic effect on SID performance when evaluating a system on mixed conditions with a single threshold. In reducing condition sensitivity, a system can be expected to act more predictably when more than one condition is encountered (such as telephone *or* microphone test samples) despite development conditions. Although methods such as meta-based calibration and trial-based calibration aim to help in this regard [11], we constrain the scope of this paper to dealing with inherent calibration of a system such that a single and simple "global" calibration becomes more applicable. To date, DNN-based SID performance has been reported in the context of constrained trial conditions. In the deployment of a system, however, the conditions to which it is applied may vary from those in which it was developed, or include a mix of trial conditions. Accordingly, we focus the remainder of this paper on the task of analyzing and reducing the sensitivity of system scores to trial conditions in order to allow the linear calibration model to operate as intended.

3. The Hybrid Alignment Framework

The two DNN-based approaches to SID described in Section 2.1 can utilize the same DNN (as illustrated in Figure 1) for either extraction of features or direct use in the process of aligning features for statistics generation. In this work, we propose a hybrid framework that uses both methods to specifically analyze the effect of BN features on the alignment process as compared to their use in statistics for the extraction of i-vectors, and the impact on SID performance[1] The motivation for this analysis comes from the inherent DNN objective to remove speaker variation to improve speaker-independent senone prediction. Though this speaker-independence is expected in the BN features, they tend to provide benefit to SID [15]. As we will show later in Section 5, their effect on calibration does not always correlate with their perceived performance benefit. Care should be taken when BN features are employed in a system.

The hybrid alignment framework relies on a UBM for fea-

[1]This work was under review when a similar framework was published in [16]. The current study differs by focusing on the calibration aspects of DNN-based speaker recognition.

ture alignment but exploits different features for the purpose of generating alignments as compared to those used in statistics calculation. This is an extension of our previous use of DNN senone posteriors γ for use in the calculation of statistics for the SID features. Specifically, instead of using the senone posteriors of a DNN trained in a supervised manner as in the DNN/i-vector approach in [1], the posteriors are calculated from a UBM trained using standard, unsupervised clustering. This UBM is trained using what we refer to as an "alignment" feature. The zero-order statistics for an audio file are given as the sum over the UBM posteriors γ from the alignment features extracted from the audio. These posteriors are then used with the corresponding SID feature, to calculate the first-order statistics. For i-vector extraction, these first-order statistics must be centralized. For this purpose, the mean and covariance is learned over a set of first-order and second-order statistics generated from the same data as used to train the UBM. One such implementation following in Figure 1 would use MFCC features as the SID features and the BN features to generate UBM posteriors for alignment of the SID features.

There are a number of anticipated benefits from this hybrid architecture. First, we are able to analyze the impact of DNN-based options for feature alignment and compare it to their impact of statistics generation for the task of SID. Secondly, in using different features for alignment compared to i-vector extraction, we can reduce computational requirements as compared to the use of concatenated features. Finally, the DNN/i-vector approach is constrained by the pre-defined number of senones used in DNN training, whereas the unsupervised clustering of bottleneck features allows for more flexibility in system design as it is not constrained in the same way.

4. Evaluation Protocol

Experiments in this work are based on the PRISM [17] and Speakers in the Wild (SITW) [18] corpora. Initial analysis experiments are conducted on PRISM using a gender-independent (GI) i-vector extractor [12] and GI classification via mixture-of probabilistic linear discriminative analysis (PLDA) models [19]. Noise-aware speech activity detection (SAD) was based on Gaussian mixture models (GMM) as previously used in [20]. All i-vectors were processed with mean and length normalization and LDA prior to PLDA [21]. In addition to the noise- and reverb-degraded audio of PRISM, we included transcoded audio files following the work in [22] for LDA and PLDA training. The test set was split in two partitions, based on speaker label, to provide both a calibration set and evaluation set. Performance metrics were evaluated on the latter.

In Section 5.3, we apply the GI system trained on PRISM data to the evaluation of the SITW dataset. This publicly available dataset contains speech from open source media from nearly 300 individuals. The audio poses significant challenges in terms of variation in real world conditions including clean interviews, red carpet interviews including babble, reverberant stadium conditions, outdoor conditions, and spontaneous noise. This variation results in severe cross-condition trials across a range of speech durations and also includes cross-gender trials. Readers are directed to [18] for more details on this database. The development partition of the SITW database was used to generate trial scores (2,597 target and 335,629 impostor trials) from which calibration parameters were learned and applied to the evaluation partition (3,658 target and 718,130 impostor trials) from which metrics were reported accordingly.

Several different features are used in this work. MFCCs of 20 dimensions were contextualized with deltas and double deltas. BN features were extracted from a 5-layer DNN consisting of 1200 nodes in each hidden layer to predict 3494 senone outputs, while the second-to-last hidden layer — the bottleneck layer — was restrained to 80 dimensions. The input features for the DNN consisted of 40 log Mel filter bank energies along with the energies from seven frames either side of a frame for a contextualized feature of 600 dimensions. The DNNs were trained by using the same dataset as used in [6]. Similarly, the input features were mean and variance normalized over the full waveform to improve channel robustness. We use a single DNN in this work for both extraction of BN and generation of senone posteriors.

Table 1: Condition-dependent and pooled-trial EER (%) / C_{llr} on the PRISM dataset from three systems with different alignment and SID feature options. Subscript int and phn denote an interview or phone call speaking style.

Condition	MFCC	BN+MFCC	DNN/iv
tel-tel	2.81 / .111	**1.20** / .070	1.23 / **.054**
tel-$\mathrm{mic}_{\mathrm{int}}$	1.25 / .084	0.81 / .072	**0.76** / **.060**
$\mathrm{mic}_{\mathrm{int}}$-$\mathrm{mic}_{\mathrm{phn}}$	2.70 / .177	1.68 / .162	**1.63** /**.118**
$\mathrm{mic}_{\mathrm{int}}$-$\mathrm{mic}_{\mathrm{int}}$	2.36 / .131	2.06 / .141	**1.87** / **.117**
pooled	3.00 / .131	2.42 / .125	**1.77** / **.099**

Two performance metrics are used in this work for both PRISM and SITW. We report Equal Error Rate (EER) to measure the discriminative power of a system and we analyze how well a system is calibrated across all operating points using a log-likelihood ratio cost metric, C_{llr} [10].

5. Results

This section analyzes the effect of different DNN-based SID approaches on calibration performance. Through analysis of different hybrid alignment framework inputs, it also provides direction into how to reduce score distribution condition-sensitivity and therefore increase the applicability of linear calibration.

5.1. Condition-dependent Analysis

First, we provide condition-dependent results on the PRISM database for which the test set was divided into two equal partitions based on speaker ID in order to perform calibration, per Section 4. Performance metrics are reported on a per-condition basis, as well as on the set of scores formed by pooling all trials. This pooled case is of particular interest for the EER metric, since a single threshold must be applied irrespective of condition, and a poor EER relative to the trial-weighted average over individual conditions is an indicator that score distributions may exhibit condition dependence. Three systems were evaluated in this manner: the MFCC system, the feature-concatenated BN+MFCC system, and the DNN/i-vector system based on MFCCs (DNN/iv).

Firstly, we focus on the telephone-only (tel-tel) EER and C_{llr} results reported in Table 1, where we can observe a gain of 37–57% in performance metrics from systems that leverage BN features or the DNN directly for alignment over the MFCC system. This finding for EER aligns with numerous studies already conducted on telephone speech and DNN-based SID [1, 15, 6]. For the alternate conditions of PRISM, the EER of the BN+MFCC improves over MFCC between 13–38%.

Table 2: Performance in terms of EER (%) / C_{llr} on the pooled trials of the PRISM dataset using different alignment and SID feature configurations in the hybrid alignment framework.

Alignment	SID Feature		
	BN	BN+MFCC	MFCC
UBM-MFCC	4.43 / .184	3.40 / .151	3.00 / .131
UBM-BN	3.66 / .163	2.51 / .126	1.90 / .101
UBM-BN+MFCC	3.33 / .154	2.42 / .125	1.85 / **.097**
DNN senones	3.16 / .152	2.42 / .132	**1.77** / .099

Table 3: Performance in terms of EER (%) / C_{llr} on the pooled trials of the SITW dataset using different alignment and SID feature configurations in the hybrid alignment framework.

Alignment	SID Feature		
	BN	BN+MFCC	MFCC
UBM-MFCC	13.81 / .435	12.22 / .392	12.30 / .414
UBM-BN	13.01 / .414	10.94 / .355	10.19 / .348
UBM-BN+MFCC	12.90 / .404	10.72 / .350	9.46 / **.312**
DNN senones	11.65 / .378	10.55 / .337	**9.26** / .321

This gain is slightly greater for the DNN/iv system[2], which may be leveraging a larger supervector space (3494 senones vs 1024 UBM components). Of interest to this study is the way that calibration trends differ significantly between the two DNN-based approaches. The BN+MFCC system is similar to the MFCC system in terms of C_{llr} for all but the tel-tel condition, while the DNN/iv system offers an 11–33% calibration improvement. This calibration benefit is strongly conveyed in the results of the "pooled" trials. Here, we can observe a 19% and 5% improvement in EER and Cllr (respectively) from BN+MFCC over MFCC; in contrast, the DNN/iv system held corresponding improvements of 41% and 24%.

When comparing systems, we can observe that the difference between the pooled trial EER and the average condition EER for the BN+MFCC system is considerably greater than the MFCC or DNN/iv system. This indicates that the score distributions from each trial condition in the BN+MFCC system are sufficiently different to prevent a single threshold from providing adequate classification in this system.

5.2. Alignment Features vs SID Features

Results in the previous section indicated a condition sensitivity in the BN+MFCC system. This sensitivity was not apparent in the MFCC or DNN/iv framework. This finding brings into question the role of each feature for the task of feature alignment versus that of first-order statistics used for i-vector extraction (i.e., the SID feature). To shed some light on the root cause of this issue, we ran systems using four different methods of feature alignment and three different SID feature configurations based on the hybrid alignment framework proposed in Section 3. Table 2 provides these comparisons in terms of EER and C_{llr}. Considering first the four methods of alignment generation, the use of DNN-based techniques for alignment is clearly an essential part of state-of-the-art technology, with MFCC-based alignments consistently providing a significant loss in performance irrespective of the choice of SID feature. When MFCCs are introduced alongside the BN features (BN+MFCC) for alignment, performance tends to improve. In contrast to the alignment feature, MFCC alone as the SID feature is optimal, and any inclusion of BN features in the first-order statistics reduces calibration and discrimination performance in pooled trials. Given these findings, we can conclude that the hybrid alignment framework provides a means to leverage the benefits of bottleneck features for alignment while providing comparable calibration and discriminative performance across pooled trail conditions to that offered by the DNN/iv framework.

[2]This finding differs from our previous publication [6] in which the DNN/iv implementation using MVN features incorrectly applied MVN based on the whole waveform, instead of just speech frames.

5.3. Speakers in the Wild (SITW) Results

Results so far have been constrained to the PRISM dataset in which discrete trial conditions were analyzed individually as well as after pooling trials. The SITW database provides an interesting scenario in which "wild" audio from open-source media is the focus. Consequently, the concept of discrete categories does not exist, but rather a range of audio-degrading artifacts at different levels. The recognition of the same speaker across such varying conditions will require that limited condition sensitivity is exhibited from the speaker recognition system applied to the task. For this task, we apply the calibration models learned from the development set of SITW to the scores of the SITW evaluation data.

Table 3 presents the results of the pooled SITW metrics. As with the observed trends on the PRISM database, the DNN-based alignment techniques improve the performance over MFCC-based alignment irrespective of the choice of the SID features. Further, the addition of MFCCs to BN features for alignment improve performance over BN features. Consistently, we can observe that the MFCC alone as the SID feature is the best choice for both PRISM and SITW. On the challenging SITW corpus, the hybrid alignment framework using BN+MFCC for alignment and MFCCs as the SID feature provides comparable performance to the DNN/i-vector approach. This amounts to over a 22% relative improvement in both metrics on the SITW corpus over a MFCC-only system, and up to a 12% gain over the traditional BN+MFCC i-vector architecture.

6. Conclusions

This article focused on the issue of calibration in current DNN-based speaker recognition algorithms and proposed a new hybrid alignment framework to address these issues. Using both PRISM and SITW databases, we demonstrated that DNN-based SID systems provide very good speaker discrimination power, when analyzed on a per-condition basis. However, the application of these systems to varying or unseen conditions poses a considerable issue with respect to calibrating at a given operating point. Through analysis of the bottleneck features in use for feature alignment and/or first-order statistics during i-vector extraction, we demonstrated that using BN features for statistics calculation stifles calibration by introducing condition sensitivity in the system score distributions. The proposed hybrid alignment framework uses BN features to generate alignments and applies these to non-BN features for first-order statistics calculation. This approach provided considerably more robust scores, with a 22% and 12% relative gain in calibration performance on PRISM and SITW, respectively, over the previous BN+MFCC i-vector architecture, and a computational advantage over prior DNN-based SID systems.

7. References

[1] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically aware deep neural network," in *Proc. ICASSP*, 2014.

[2] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," in *Proc. Interspeech*, 2015.

[3] D. Garcia-Romero and A. McCree, "Insights into deep neural networks for speaker recognition," in *Proc. Interspeech*, 2015.

[4] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Speaker Odyssey*, 2014.

[5] Y. Lei, L. Ferrer, M. McLaren, and N. Scheffer, "A deep neural network speaker verification system targeting microphone speech," in *in Proc. Interspeech*, 2014.

[6] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. IEEE ICASSP*, 2015.

[7] P. Matejka, L. Zhang, T. Ng, S. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in *Proc. Speaker Odyssey*, 2014.

[8] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," in *Proc. Speaker Odyssey*, 2014.

[9] L. Ferrer, Y. Lei, and M. McLaren, "Study of senone-based deep neural network approaches for spoken language recognition," *Submitted to IEEE Trans. Audio Speech and Language Processing*, 2015.

[10] N. Brummer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

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

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Speech and Audio Processing*, vol. 19, pp. 788–798, 2011.

[13] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[14] M. McLaren, N. Scheffer, L. Ferrer, and Y. Lei, "Effective use of DCTs for contextualizing features for speaker recognition," in *Proc. ICASSP*, 2014.

[15] M. McLaren, L. Ferrer, and A. Lawson, "Exploring the role of phonetic bottleneck features for speaker and language recognition," in *Proc. IEEE ICASSP*, 2016.

[16] P. Matejka, O. Glembek, O. Novotny, O. Plchot, F. Grezl, L. Burget, and J. Cernocky, "Analysis of DNN approaches to speaker identification," in *Proc. IEEE ICASSP*, 2016.

[17] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and S. N., "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," in *Proc. NIST 2011 Workshop*, 2011.

[18] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Submitted to Interspeech*, 2016.

[19] M. Senoussaoui, P. Kenny, N. Brummer, E. De Villiers, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender independent speaker recognition," in *Proc. Int. Conf. on Speech Communication and Technology*, 2011.

[20] 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.

[21] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[22] M. McLaren, V. Abrash, M. Graciarena, Y. Lei, and J. Pesn, "Improving robustness to compressed speech in speaker recognition," in *Proc. Interspeech*, 2013, pp. 3698–3702.