Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Source Normalization for Language-Independent Speaker Recognition using i-vectors

Mitchell McLaren, Miranti Indar Mandasari and David A. van Leeuwen

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
{m.mclaren, m.mandasari, d.vanleeuwen}@let.ru.nl

Abstract

Source-normalization (SN) is an effective means of improving the robustness of i-vector-based speaker recognition for under-resourced and unseen cross-speech-source evaluation conditions. The technique of source-normalization estimates directions of undesired within-speaker variation more accurately than traditional methods when cross-source variation is not explicitly observed from each speaker in system development data. Source normalization can be incorporated into Within Class Covariance Normalization (WCCN) as an effective preprocessing step to Probabilistic Linear Discriminant Analysis (PLDA) based speaker recognition with i-vectors.

This paper proposes to extend the application of source-normalization to the reduction of language-dependence in PLDA speaker recognition by normalising for the variation that separates languages. Evaluated on the NIST 2008 and 2010 speaker recognition evaluation (SRE) data sets, the proposed Language Normalized WCCN (LN-WCCN) provides relative improvements of 26% in minimum DCF and 14% in EER under multilingual scenarios without detriment to common English-only conditions. LN-WCCN is also shown to significantly improve calibration performance when calibration parameters are learned from scores mismatched to evaluation conditions.

1. Introduction

Speaker recognition technology based on i-vectors currently dominates the research field due to its state-of-the-art performance, low computational cost and the suitability of i-vectors for use with many existing pattern recognition techniques such as Linear Discriminant Analysis (LDA), Probabilistic LDA (PLDA) and Support Vector Machines (SVM) [1, 2]. I-vectors represent a speech utterance as a fixed length vector extracted from a low-dimensional subspace that bounds all sources of variability. Consequently, i-vectors represent both between- and within-speaker variation; the later of which is detrimental to system performance. Session compensation is therefore required to minimise the impact of within-speaker variation.

Session compensation aims to improve system robustness to differences between utterances of the same speaker. Encompassed by the term session variation, these differences are induced by factors such as speech source, transmission channels, background noise, and speaker characteristics including health, age, accent and language. In order to robustly compensate for these factors, system development data must be representative of the variability in such characteristics [3]. Unfortunately, the amount of data available for this purpose is often limited thereby restricting the benefit from traditional session compensation techniques such as Within-class Covariance Normalization (WCCN), LDA, and PLDA [4, 5]. Common to each of these approaches is the need to estimate a within-class covariance or scatter matrix which typically does not capture all sources of variation detrimental to system performance [4].

Source Normalization (SN) was recently proposed to accommodate the shortcomings of traditional scatter estimation in the context of development datasets commonplace in the speaker recognition community [4, 6]. Specifically, development data sets do not typically consist of a minimum of one utterance from every 'source' of variation from each speaker needed to accurately estimate within-class scatter. Sources of variation are labelled dataset characteristics that may contribute to differences between utterances of the same person as observed by the system. Source normalization was originally developed to compensate for speech source variation (i.e., the differences between microphone and telephone sourced speech) and offers significant improvements in cross speech-source trials and under-resourced microphone speech conditions [4]. One source of variation that has received limited focus since the invention of i-vectors is that of language.

Speaker recognition systems are commonly tailored toward English speech due to the ample resources available for system development and the focus of the recent NIST 2010 Speaker Recognition Evaluation (SRE) [7]. Consequently, both discrimination and calibration performance is difficult to maintain under multilingual trial conditions [8, 9, 10]. This is largely due to the relatively limited development data available for non-English languages and the fact that each speaker in the data set does not speak every language, thus resulting in a suboptimal estimate of the within-speaker covariance to be suppressed. It was precisely this type of scenario for which source normalization was developed.

In this work, we extend source normalization to improve the robustness of multilingual PLDA-based speaker recognition using i-vectors. The approach taken involves subjecting i-vectors to SN-WCCN prior to PLDA modelling where the language is the source of variation to be suppressed. Thus, we evaluate Language Normalized WCCN (LN-WCCN) on the all-language telephony trials of SRE'08 as well as the English-only trials of both SRE'08 and SRE'10 in order to observe the effect of language normalization on the well-developed English-only condition. Further, the effect of LN-WCCN on score calibration is analysed in the context of matched and mismatched calibration and evaluation data.

This article provides an overview of speaker recognition using i-vectors in Section 2. Source normalization and the proposed LN-WCCN are detailed in Section 3. Section 4 describes the experimental protocol used in this study followed by experimental results and analysis in Section 5.

This research was funded by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant no. 238803.

2. Speaker recognition using i-vectors

In recent years, speaker recognition research has focussed on the use of i-vectors to represent an utterance [1, 8]. Once obtained, i-vectors can be used in conjunction with many straightforward pattern recognition techniques such as LDA and cosine distance scoring to obtain a high level of performance. Current state-of-the-art technology consists of several preprocessing stages followed by PLDA modelling of i-vectors [2, 11]. This section describes the PLDA framework utilised in this work.

2.1. I-vector extraction

I-vectors can be viewed as a compact representation of a speech utterance extracted from a low-dimensional subspace, $\boldsymbol{T}$, that bounds the main directions of between-utterance variability. Referred to as the total variability subspace, this subspace is estimated from a large set of development data via factor analysis [12]. An i-vector is the latent factor vector, $\boldsymbol{w}$, obtained from the Gaussian Mixture Model (GMM) representation,

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}, \tag{1}$$

where $\boldsymbol{m}$ is the speaker- and session-independent mean supervector taken from a Universal Background Model (UBM). The low-rank i-vector (400 dimensions in this work) has a standard normal distribution $\mathcal{N}(0, 1)$ and is given as the maximum-a-posteriori point estimate in the space defined by $\boldsymbol{T}$ based on the observed Baum-Welch statistics from an utterance. Further details on subspace training and i-vector extraction can be found in [12] and [8].

2.2. I-vector preprocessing

Traditional PLDA can be utilised in conjunction with a number of simple i-vector preprocessing stages to achieve state-of-the-art performance comparable to the more complex heavy-tailed PLDA model [13, 2, 11]. These processes include WCCN [14], followed by the normalization of i-vector length [11].

2.2.1. Within-Class Covariance Normalization (WCCN)

As the name suggests, Within-Class Covariance Normalization (WCCN) [14] normalizes the within-speaker covariance of the i-vector space. This process prevents the subsequently trained PLDA model from being biased toward directions of relatively high variation. Fundamental to WCCN is the estimate of within-speaker scatter,

$$\boldsymbol{S}_W = \sum_{s=1}^{S} \sum_{n=1}^{N_s} (\boldsymbol{w}_n^s - \boldsymbol{\mu}_s)(\boldsymbol{w}_n^s - \boldsymbol{\mu}_s)^t, \tag{2}$$

where S is the number of speakers in the development dataset, each of whom have N_s utterances, and $\boldsymbol{\mu}_s = \frac{1}{N_s} \sum_{n=1}^{N_s} \boldsymbol{w}_n^s$ is the mean of the i-vectors from speaker s. Given an estimate of $\boldsymbol{S}_W$, the WCCN transform $\boldsymbol{B}$ is found through the Cholesky decomposition of $(\frac{1}{S}\boldsymbol{S}_W)^{-1} = \boldsymbol{B}\boldsymbol{B}^t$.

2.2.2. Length normalization

Normalising i-vectors to have a unit length has been shown to greatly improve PLDA-based performance [2, 11]. Garcia-Romero *et al.* [11] found that differences in i-vector length between system development and evaluation data contributes to system errors and can be counteracted through length normalization. This straightforward process better fits the distribution of i-vectors to the Gaussian assumptions made by the PLDA model, due to the fact that an D-dimensional Gaussian for high D has most of its probability density around a thin shell at constant distance from the mean. A raw i-vector is thus preprocessed using WCCN and length-normalization via

$$\overline{\boldsymbol{w}} = \frac{\boldsymbol{B}^t \boldsymbol{w}}{|\boldsymbol{B}^t \boldsymbol{w}|}. \tag{3}$$

2.3. Probabilistic Linear Discriminant Analysis (PLDA)

Speaker detection involves comparing two speech utterances to determine whether or not they were uttered by the same speaker. Probabilistic Linear Discriminant Analysis (PLDA) is a probabilistic approach that provides this information directly in terms of a likelihood ratio R. The comparison of two i-vectors $\overline{\boldsymbol{w}}_1$ and $\overline{\boldsymbol{w}}_2$ is given by the ratio of hypothesis H_{tar} that both i-vectors originate from the same speaker and H_{non} that they were uttered by different speakers. This can be formulated as

$$R = \frac{P(\overline{\boldsymbol{w}}_1, \overline{\boldsymbol{w}}_2 | H_{\text{tar}})}{P(\overline{\boldsymbol{w}}_1 | H_{\text{non}}) P(\overline{\boldsymbol{w}}_2 | H_{\text{non}})}, \tag{4}$$

where $P(\overline{\boldsymbol{w}}_1, \overline{\boldsymbol{w}}_2 | H_{\text{tar}})$ and $P(\overline{\boldsymbol{w}} | H_{\text{non}})$ can be determined from a trained PLDA model following the strategy taken in [2].

The PLDA model assumes that i-vector $\overline{\boldsymbol{w}}_s$ can be modelled as,

$$\overline{\boldsymbol{w}}_s = \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{\epsilon} \tag{5}$$

where $\boldsymbol{V}$ and $\boldsymbol{U}$ are subspaces that bound the major directions of speaker and session variation, respectively. Loading factors $\boldsymbol{y}_s$ and $\boldsymbol{x}$ have a standard normal distribution and $\boldsymbol{\epsilon}$ represents the residual variation with diagonal covariance matrix. In this work, $\overline{\boldsymbol{w}}_s$ has 400 dimensions while subspaces $\boldsymbol{V}$ and $\boldsymbol{U}$ are tuned to between 50–100 dimensions as detailed in Section 4.

3. Language-Normalization

This section extends SN-WCCN to the task of suppressing the variation that separates languages in the i-vector space with the objective of improving system robustness to multiple languages.

3.1. Source normalization

Source Normalization (SN) [6, 4] was initially developed to improve recognition performance in cross speech source conditions and under-resourced trial conditions through obtaining a more accurate estimation of the scatter matrices used during LDA and WCCN optimisation in the context of a suboptimal dataset. The use of a suboptimal dataset for scatter estimation is commonplace in speaker recognition research and can be defined as one in which every speaker does not provide an utterance from every source. A 'source' of variation is typically a data-labelled characteristic such as transducer type (i.e., microphone or telephone) that contributes to differences in i-vectors extracted from the same speaker and, subsequently, to a degradation in speaker recognition performance. In this context, the use of (2) to estimate within-speaker scatter for the purpose of WCCN does not adequately represent the directions of variation within speakers and WCCN is therefore unable to properly normalize for this variation [4].

Source normalization is based on the fact that the total variation of i-vectors $\boldsymbol{S}_T = \boldsymbol{S}_B + \boldsymbol{S}_W$, where the total scatter matrix $\boldsymbol{S}_T = \sum_{n=1}^{N} \overline{\boldsymbol{w}}_n \overline{\boldsymbol{w}}_n^{t\,1}$. Determining $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ is thus a breakdown in total variation. It was demonstrated in [4] that the

[1]The center of the i-vector space is a null vector due to the factor analysis assumption and therefore not used to calculate $\boldsymbol{S}_T$.

traditional method of scatter estimation can result in a between-speaker scatter affected source variation — variation that is, in fact, within-speaker variation. Source normalization is a two stage process in which a normalized between-speaker scatter, $\hat{\boldsymbol{S}}_B$, is estimated to be void of source variation after which the within-speaker scatter is found as the residual variation,

$$\hat{\boldsymbol{S}}_W = \boldsymbol{S}_T - \hat{\boldsymbol{S}}_B. \quad (6)$$

Fundamental to SN is the estimation of scatter $\hat{\boldsymbol{S}}_B$ which is an accumulation of source-dependent scatter matrices $\boldsymbol{S}_B^{\text{src}}$ for each source of interest, src. This can be formulated as

$$\hat{\boldsymbol{S}}_B = \sum_{\text{src}} \boldsymbol{S}_B^{\text{src}}, \quad (7)$$

$$\boldsymbol{S}_B^{\text{src}} = \sum_{s=1}^{S_{\text{src}}} N_s^{\text{src}} (\boldsymbol{\mu}_s^{\text{src}} - \boldsymbol{\mu}_{\text{src}})(\boldsymbol{\mu}_s^{\text{src}} - \boldsymbol{\mu}_{\text{src}})^t. \quad (8)$$

where S_{src} is the number of speakers with i-vectors from source src in the development dataset and $\boldsymbol{\mu}_s^{\text{src}}$ is the mean of the N_s^{src} i-vectors from speaker s and source src. Key to the approach is fixing the the center of the i-vector space to the source mean $\boldsymbol{\mu}_{\text{src}} = \frac{1}{N_{\text{src}}} \sum_{n=1}^{N_{\text{src}}} \overline{\boldsymbol{w}}_n^{\text{src}}$ where N_{src} is the number of i-vectors available for source src.

3.2. Source-Normalized WCCN (SN-WCCN)

Source Normalized WCCN (SN-WCCN) can be implemented by using the normalized within-speaker scatter $\hat{\boldsymbol{S}}_W$ from (6) in place of $\boldsymbol{S}_W$ in (2). That is, decomposing $(\frac{1}{S}\hat{\boldsymbol{S}}_W)^{-1} = \boldsymbol{B}\boldsymbol{B}^t$. In [4], SN-WCCN was shown to be comparable to SN-LDA where the latter additionally exploits the information of the normalized between-class scatter. Utilised for i-vector preprocessing, SN-WCCN was recently integrated into the PLDA-based i-vector framework in the context of gender-independent speaker recognition in which the source to be normalized was both gender and speech source [5]. In this manner, source variation that dominates the i-vector space [4] is heavily attenuated to allow the subsequent PLDA modelling process to better exploit true speaker discriminative information.

3.3. Language-Normalized WCCN (LN-WCCN)

In this study, we focus on the scenario of multilingual speaker recognition made particularly difficult due to speakers being able to speak in multiple languages. For example, it is not trivial to recognize whether two speech samples of different languages were uttered by the same speaker due the significant differences in phonetic variability. As illustrated later in Section 5.1, each language is represented differently in i-vector space, leading to the surmise that language is a source of undesirable variation. As speakers in the development dataset can not provide utterances in every language of interest, a within-speaker scatter matrix estimated via source normalization is required to reduce the effect of language variation.

In this work, we extend SN-WCCN to the task of removing variation that separates languages via *Language-Normalized* WCCN (LN-WCCN). This can be implemented by labelling language as the source for SN-WCCN. Based on previous research findings [4], LN-WCCN is expected to improve system performance in both cross-language trials and under-resourced non-English language conditions. Noteworthy is that LN-WCCN requires language-labels only to estimate of the within-speaker scatter during system development and is language-blind at trial time.

	Male		Female	
Language	#Spkr	#Seg	#Spkr	#Seg
Arabic	108	740	82	517
English	3271	21697	3958	27376
Farsi	59	238	60	263
French	54	242	62	278
German	55	255	61	283
Hindi	73	307	50	218
Japanese	57	244	61	266
Korean	52	226	62	267
Mandarin	194	1073	280	1423
Russian	27	199	47	416
Spanish	118	661	172	967
Tamil	63	220	51	216
Thai	–	–	2	6
Urdu	–	–	2	5
Vietnamese	55	240	62	263
Yue	2	2	–	–

Table 1: Multilingual resources used for system development.

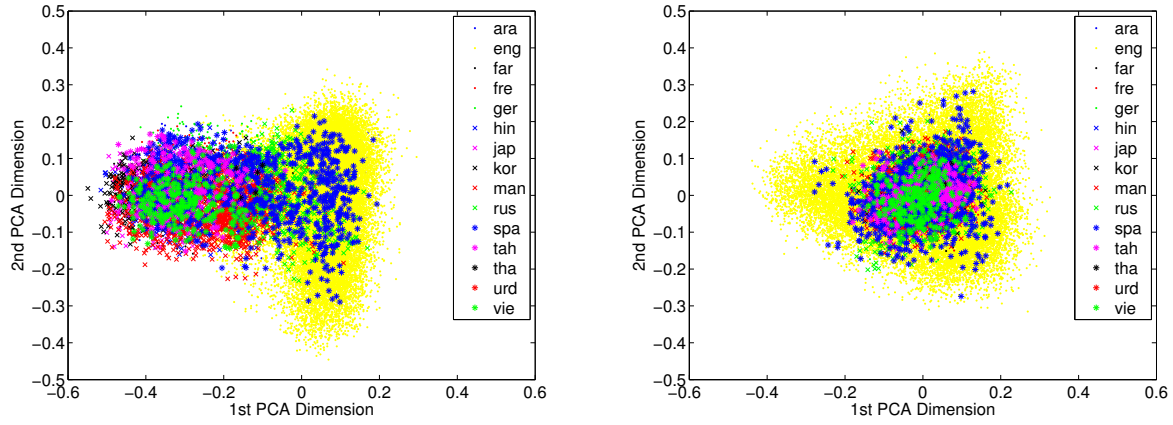
4. Experimental Protocol

The recent NIST 2008 and 2010 SRE corpora are used to evaluate the proposed approach of language normalization. Results are reported for telephony (*tel-tel*) multilingual trials (det6) and English-only trials (det7) on the SRE'08 database and the telephony English-only trials from det5 of the SRE'10 extended protocol. Performance was evaluated using the equal error rate (EER) and a normalized minimum decision cost function (DCF) calculated using effective prior odds of 1/9.9 and 1/999 for SRE'08 and SRE'10, respectively [15]. In all approaches, the number of PLDA subspace dimensions was evaluated in steps of 50 in order to minimise the average of $(\text{DCF} + 10 \times \text{EER})$ from det5 of SRE'10. The resulting dimensions were 100 speaker and 100 session dimensions when preprocessing i-vectors with WCCN and 100 speaker and 50 session dimensions for LN-WCCN. These optimised subspaces were then used in the evaluation of SRE'08.

Speech activity detection was performed as in [16]. Gender-dependent, 2048-component UBMs were trained on 20-dimensional, feature-warped MFCCs (including C_0) with deltas and double-deltas appended. UBM training data included telephone and microphone speech sourced from the NIST 2004—2006 SRE corpora and LDC releases of Fisher English, Switchboard II: phase 3 and Switchboard Cellular (parts 1 and 2). The total variability subspace of 400 dimensions was trained using the same data along with data from Switchboard I, Switchboard II: phase 1 and the Callfriend database. The WCCN and LN-WCCN transforms and PLDA models were trained using the same data as used for subspace training but limited to telephone speech.

4.1. Multilingual development speech

System development data includes speech from a variety of languages represented by the aforementioned NIST SRE (2004–2006), Fisher and Switchboard series of corpora along with the Callfriend database commonly used for language identification. The Callfriend database consists of speech from 12 languages recorded as 30 minute conversations from 120 speakers of each language. Each 30 minute conversation side was split into multiple segments containing 5 minutes of audio to match the length of NIST telephone conversations and to provide a limited es-



(a) WCCN (b) LN-WCCN

Figure 1: Projection of language-labelled i-vectors into 2D PCA space after applying (a) WCCN or (b) LN-WCCN. This represents the distribution of languages in i-vector space prior to training the PLDA classifier.

timate of within-speaker variation for the Callfriend speakers. This resulted in approximately 400–1000 utterances per language from Callfriend which included Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. Additional languages represented in the SRE 2004–2006 datasets include Russian, Yue, Thai, and Urdu. Table 1 details the number of speakers and segments available for each represented language on a per gender basis.

4.2. Score Calibration Protocol

The calibration process was conducted using the FoCal [17] toolkit which is based on a linear transformation,

$$\ell = \alpha_0 + \alpha_1 s \tag{9}$$

to produced calibrated log-likelihood-ratio (LLR) ℓ from raw scores s. The linear transformation weighting parameters α_0 and α_1 were trained from a set of development scores via logistic regression [10]. This transformation was optimized for the NIST SRE prior 1/9.9 [15]. In this paper, the calibration performance is evaluated in an application independent approach based on [18], and presented in terms of cost of LLR (C_{llr}), the minimum cost of LLR (C_{llr}^{min}), and miscalibration $C_{mis} = C_{llr} - C_{llr}^{min}$. The calibration experiments were carried out in a gender-dependent manner using SRE'10 for development data, while SRE'08 was specified as the evaluation dataset in order to provide unseen language variability in the calibration evaluation.

5. Results

The following experiments illustrate the effectiveness of the proposed language-normalized WCCN to improve the robustness of PLDA-based speaker recognition. As detailed in Section 4, tuning of system parameters including PLDA subspace dimensions was done entirely on SRE'10 so as to allow for an unbiased comparison of SRE'08 results. This is of particular interest since SRE'10 is an English-only data set.

5.1. Analysis of language-normalized i-vector space

Prior to analysing performance trends, we have often found it beneficial to observe how a given technique changes the way in which data lies in the i-vector space. In a similar manner to previous publications [4, 5], i-vectors from the female PLDA training dataset were first processed using either WCCN or LN-WCCN. Each set of i-vectors was used to train a corresponding 2D PCA space into which the same i-vectors were projected. This final projections are depicted in Figure 1 and labelled based on the language to which i-vectors correspond. Figure 1(a) represents i-vectors processed with WCCN. The distribution of i-vectors from other languages (darker scatter) is noticeably disjoint from i-vectors corresponding to English speech (yellow scatter). It is the objective of LN-WCCN to suppress this variation. Figure 1(b) illustrates that LN-WCCN successfully removed the variation attributed to language differences and resulted in a less skewed distribution of i-vectors with a common center. Based on these observations, it is expected that using LN-WCCN instead of WCCN will provide added robustness to cross-language speaker comparisons in a subsequently trained PLDA model.

Corpus	Trial Lang.	WCCN Min. DCF	WCCN EER	LN-WCCN Min. DCF	LN-WCCN EER
SRE'08	All	.0329	5.98%	**.0244**	**5.12%**
SRE'08	English	.0111	**2.28%**	.0111	2.36%
SRE'10	English	**.4510**	**2.71%**	.4540	2.85%

Table 2: SRE'08 det6 (All) and det7 (English) and SRE'10 det5 extended (English) results comparing WCCN to Language Normalized-WCCN across different trial conditions. PLDA subspaces were tuned on SRE'10 results.

5.2. Language normalization

This section compares PLDA performance when modelling i-vectors preprocessed with WCCN or the proposed LN-WCCN using the NIST SRE'08 and SRE'10 databases. While focus is given to the multilingual SRE'08 corpus, the SRE'10 det5 (extended) protocol was evaluated to observe the effect of LN-WCCN on the English-only trials of this recent data set.

Table 2 summarizes results from all-language trials (det6) and English-only trials (det7) on the SRE'08 database. In the all-language trials (top line of the table), LN-WCCN was found to provide a relative improvement of 26% in minimum DCF and 14% in EER over the use of WCCN. This demonstrates that suppression of language related variation from i-vectors provides improved robustness to multilingual speech in a PLDA-based speaker recognition system. Interestingly, the English-only trials (det6) of SRE'08 were largely unaffected by the introduction

Languages	Trial Restriction	# Trials	WCCN Min. DCF	WCCN EER	LN-WCCN Min. DCF	LN-WCCN EER
All	None	35896	.0329	5.98%	**.0244**	**5.12%**
All	Different Language	13967	.0378	7.65%	**.0309**	**6.39%**
non-English	None	4853	.0427	7.95%	**.0338**	**6.84%**
non-English	Same Language	4168	.0420	7.46%	**.0332**	**6.86%**
non-English	Different Language	685	.0279	**6.69%**	**.0248**	8.51%

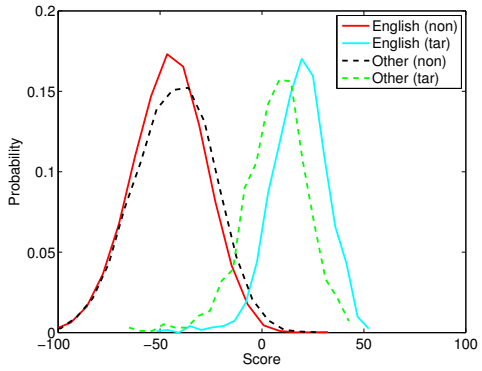
Table 3: Comparison of performance from WCCN and Language-Normalized WCCN on language-conditioned subsets of the SRE'08 det6 trials.

of language normalization into the system. LN-WCCN appears, therefore, to hold the desirable characteristic of not 'trading-off' between performance of the targeted multilingual conditions and the English-only trials. This was further explored by evaluating the English-only telephone trial det5 protocol (extended) from SRE'10 using both WCCN and LN-WCCN preprocessing steps. It was this condition for which the PLDA subspace dimensions were tuned. The last row in Table 2 shows that the difference between LN-WCCN and WCCN on the SRE'10 trials was minimal. This supports the conclusion that removing the variation that separates languages from i-vector space does not degrade the representation of English spoken speech.

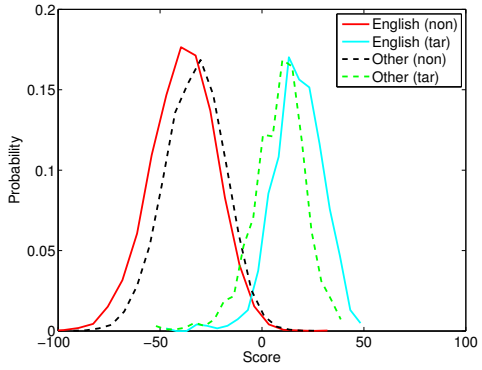
Results from the SRE'08 all-language condition (det6) were divided into subsets in order to observe where LN-WCCN provided most benefit. Subsets of results included different language trials (det7\det6), and three conditions involving only non-English speech: same language, different language and all non-English trials. Table 3 details the system performance from these subsets when using WCCN or LN-WCCN along with the number of trials in each set. The top line of the table references the performance of det6 trials from SRE'08. It can be observed that, with the exception of the last row in the table, the remaining subsets found similar improvements of 18–21% in minimum DCF and 8–16% in EER when using LN-WCCN relative to WCCN. The non-English, different language trials subset (last row in table) lacks the number of trials needed to provide meaningful results but was included for completeness. From the findings in this section, we can conclude that in the context of a system developed using a majority of English spoken speech, language normalization offers robustness when encountering both cross-language trials as well as under-resourced languages without degradation to commonly targeted English-only conditions.

5.3. Score distributions

The previous section illustrated that LN-WCCN provided robustness to the multilingual trials (det6) of SRE'08. In this section we analyse the distributions of the English-only and other-language trial scores to shed some light on why LN-WCCN is more effective than WCCN in the multilingual context. The SRE'08 target and non-target score distributions from English trials (det7) and other trials (det7\det6) were thus plotted for WCCN in Figure 2(a) and for LN-WCCN in Figure 2(b). It can be observed that in both plots, the English and other score distributions have different characteristics, particularly in terms of mean score. It can be seen, however, that the way in which the other language score distributions shift relative to English scores differs between WCCN and LN-WCCN. In the case of WCCN, the non-target score distribution is similar between English and other language trials, while the target score distribution for English trials shows a considerable positive shift. In contrast, LN-WCCN offered a global compression of other language trial scores around a central point. This, in turn, allowed



(a) WCCN



(b) LN-WCCN

Figure 2: Score distributions illustrating the differences between English-only trials and other language trials of SRE'08 det6 for PLDA pre-processed with (a) WCCN or (b) LN-WCCN.

for the more compatible pooling of the English and other language score distributions.

It can also be observed that the ratio of non-target score variance to target score variance was lower in the case of LN-WCCN. As observed in Figure 3, this has the effect of rotating the corresponding detection error trade-off (DET) curve anti-clockwise [19], thus explaining the greater relative improvement in DCF compared to EER observed in Section 5.2 when introducing language-normalization into the system.

5.4. Calibration Performance

In the context of the NIST SRE, unseen evaluation data often exhibits characteristics (e.g. language spoken or microphone type) which have never been seen in previous evaluations. This leads to a calibration problem due to the information in the development stage of calibration not being completely representative of the unseen evaluation data. As discussed in Section 5.2, language normalization was shown to offer robustness in cross-

Calibration Condition (Language)	Pre-processing	Male C_{llr}	Male C_{llr}^{min}	Male C_{mis}	Female C_{llr}	Female C_{llr}^{min}	Female C_{mis}
Mismatched (All)	WCCN	.244	.190	.054	.333	.256	.077
	LN-WCCN	**.169**	**.143**	**.026**	**.267**	**.217**	**.050**
Matched (English)	WCCN	**.084**	**.070**	.014	**.140**	**.110**	**.030**
	LN-WCCN	.088	.076	**.012**	.146	.114	.032

Table 4: SRE'08 actual cost of LLR (C_{llr}), minimum cost of LLR (C_{llr}^{min}), and mis-calibration ($C_{mis} = C_{llr} - C_{llr}^{min}$) for PLDA systems with WCCN or LN-WCCN pre-processing. The calibration parameters were trained on the English-only scores from SRE'10.

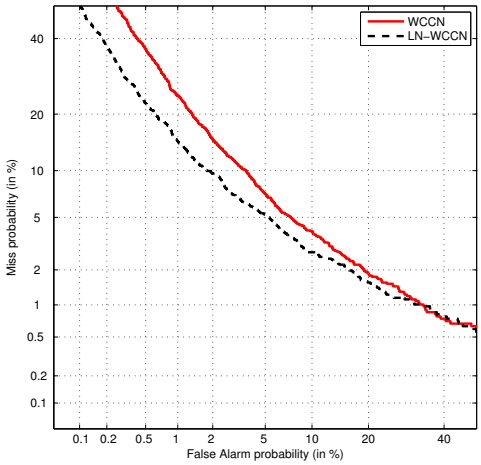


Figure 3: Plot of DET curves for SRE'08 det6 (All-language) scores comparing WCCN and LN-WCCN for i-vector preprocessing.

language and under-resourced language trials from a system tuned on English speech. We explore whether this beneficial trend of language normalization extends to the system calibration performance by evaluating two different calibration conditions — *matched* and *mismatched*. The calibration parameters were learned on the *English-only* SRE'10 det5 dataset, and the calibration was evaluated on SRE'08 dataset for both matched (det7, *English-only*) and mismatched (det6, *All-languages*) conditions.

Table 4 details the calibration performance of WCCN and LN-WCCN systems which were evaluated on the matched and mismatched calibration conditions. In the mismatched condition, all calibration metrics of LN-WCCN were reduced by 15% to 51% relative to the WCCN metrics. The largest of these relative improvements were 51% and 35% in terms of miscalibration (C_{mis}) in the male and female trials, respectively. In the matched condition, a different trend was found compared to the mismatched condition, where relatively small differences were observed between the WCCN and LN-WCCN calibration metrics. Figure 4 summarizes the effect of language normalization on calibration performance by depicting the miscalibration (C_{mis}) of matched and mismatched conditions. The figure illustrates that when compared to WCCN, LN-WCCN successfully reduced miscalibration metrics in the mismatched condition, while offering a comparable level of miscalibration in the matched condition. These results demonstrate the ability of language normalization to improve system calibration performance in the context of mismatched language conditions.

6. Conclusion

This work extended the application of source normalization to the task of improving the language-independence of state-of-

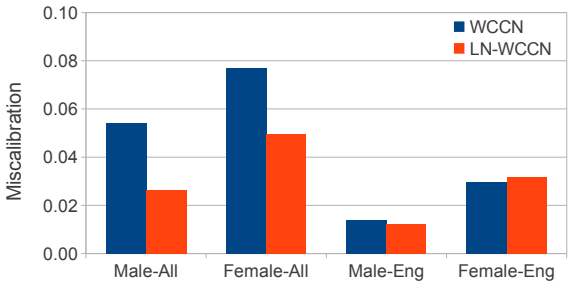


Figure 4: Comparing the miscalibration of WCCN and LN-WCCN (lower is better). Calibration parameters learned on SRE'10 det5-English and evaluated on SRE'08 det6-All (mismatched) and det7-English (matched) conditions.

the-art PLDA-based speaker recognition with i-vectors. Using language labels to identify different sources in the development data set, language normalized WCCN (LN-WCCN) was proposed as an i-vector preprocessing stage. Evaluated on the multilingual telephony conditions of SRE'08, LN-WCCN provided improvements of 26% in minimum DCF and 14% in EER relative to WCCN. These improvements were achieved without detriment to the commonly targeted English-only trial conditions or SRE'08 and SRE'10. Additionally, LN-WCCN was found to improve calibration performance when development and evaluation data were mismatched with respect to language.

7. References

[1] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. Interspeech*, 2009, pp. 1559–1562.

[2] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proc. IEEE ICASSP*, 2011.

[3] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey Speaker and Language Recogntion Workshop*, 2010.

[4] M. McLaren and D.A. van Leeuwen, "Source-normalised LDA for robust speaker recognition using i-vectors," *IEEE Trans. Audio Speech and Language Processing*, vol. 20, pp. 755–766, 2011.

[5] M. McLaren and D.A. van Leeuwen, "Gender-independent speaker recognition using source normalisation," in *accepted into Interspeech*, 2012.

[6] M. McLaren and D.A. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Proc. IEEE ICASSP*, 2011, pp. 5456–5459.

[7] National Institute of Standards and Technology, *NIST Speaker Recognition Evaluation site*, Available: http://www.itl.nist.gov/iad/mig/tests/sre/.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE*

Trans. Audio, Speech and Language Processing, vol. 19, pp. 788–798, 2011.

[9] S.S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, "The SRI NIST 2008 speaker recognition evaluation system," in *Proc. IEEE ICASSP*, 2009, pp. 4205–4208.

[10] N. Brummer, L. Burget, J.H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

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

[12] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, , and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, pp. 980–988, 2008.

[13] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010.

[14] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Ninth Int. Conf. on Spoken Language Processing*, 2006, pp. 1471–1474.

[15] David A. van Leeuwen, Alvin F. Martin, Mark A. Przybocki, and Jos S. Bouten, "NIST and TNO-NFI evaluations of automatic speaker recognition," *Computer Speech and Language*, vol. 20, pp. 128–158, 2006.

[16] M. McLaren and D.A. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *Proc. NIST SRE Workshop*, 2011.

[17] Niko Brümmer, *FoCal-II: Toolkit for calibration of multi-class recognition scores*, August 2006, Software available at http://www.dsp.sun.ac.za/∼nbrummer/focal/index.htm.

[18] D. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," *Speaker Classification I*, pp. 330–353, 2007.

[19] 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.