CSLT TECHNICAL REPORT-20160019 [Saturday 7th January, 2017]

# Language-aware PLDA for Multilingual Speaker Recognition

Askar Rozi[1,2]
, Dong Wang[1,3]
, Lantian Li[1,2]
and Thomas Fang Zheng[1,3*]

*Correspondence:
fzheng@tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

**Abstract**

Multilingual speaker recognition involves multilingual speech data in model training, which empowers the system to handle recognition requests in multiple languages. The multilingual training approach augments data from multiple languages, but inevitably introduces probability dispersion, due to the more complex language conditions. This paper proposes a language-aware training approach for PLDA which involves language information when training the PLDA model.

The proposed approach has been evaluated with the i-vector/PLDA framework using the CSLT-CUDGT2014 Chinese-Uyghur bilingual speech database. The experimental results show that the language PLDA training resulted in a relative EER reduction of 15.38% in the Chinese test and 20.07% in the Uyghur test.

**Keywords:** Multilingual speaker recognition; i-vector; PLDA

## 1 Introduction

Speaker recognition accepts or rejects a claimed identity of a speaker based on speech input. After decades of research, performance of speaker recognition systems is fairly good if the training data are sufficient and the test condition matches the condition of the training data [1, 2]. However, if there is mismatch between the conditions of training and test speech, performance of speaker recognition systems is usually degraded significantly [3]. The mismatch may be on recording device, background noise, speaking style, and others.

Among these mismatches, language variability is a special type. On one hand, it is widely assumed that speaker recognition is language independent, because speaker traits are mostly determined by acoustic features. It is not our experience that a speaker cannot be recognized when she/he speaks a different language. On the other hand, many studies have confirmed that language mismatch indeed leads to serious performance degradation for speaker recognition systems [4, 5]. There are two types of language mismatches in speaker recognition. The first type is the mismatch between enrollment and test, i.e., enroll in one language while test in another. The second type is the mismatch between model training and system operating, i.e., system is trained in one language but operated (enrollment plus test) in another language. The first mismatch is often encountered when the users are mixlingual, i.e., they use multiple languages in their daily life. The second mismatch is often encountered when the system is migrated from one language to another.

A multitude of research has been conducted to deal with language mismatch, either the one between enrollment and test or the one between training and operating. For example, Ma [4] studied the enrollment-test mismatch and found that it causes significant performance degradation for speaker recognition. Auckenthaler [5] investigated the mismatch between training and operating, within the popular universal background model-Gaussian mixture model (UBM-GMM) architecture [6]. They found considerable performance degradations if the speech data used to training the UBM and the speech data used to enroll/test speakers are in different languages. Abhinav [7] studied the same problem within the state-of-the-art i-vector architecture [8], and investigated both the enrollment-test mismatch and the training-operating mismatch. Their results confirmed that language mismatch, in spite where it occurs, leads to significant performance degradation. These results mentioned above seem opposite to our intuition that speaker traits are independent of language. We attribute the discrepancy between the empirical results and the intuition to the 'engineering' part of the recognition system: it is the models (e.g., UBMs, i-vectors, speaker GMMs) rather than the speaker characteristics that are language dependent. These models are trained in one language and are not well suited to other languages.

A simple yet effective approach to deal with language mismatch is multilingual training. This approach employs multilingual data to train the system so that all the languages in both enrollment and test are covered. All the studies mentioned above confirmed that multilingual training can largely recover the performance degradation caused by language mismatch. Particularly, the experiments presented in [7] demonstrated that even with a small amount of data from the target language, the system can obtain significant performance gains in the multilingual environment.

From the perspective of model training, the effect of the multilingual training is two-fold: on one hand, it involves more training data and therefore tends to generate stronger models; on the other hand, the model covers multiple languages and therefore the probability distribution is less concentrated compared to monolingual models. An ideal multilingual training should keep the advantage in enriched data but alleviates the effect of language mixing. This paper proposes a language-aware multilingual training approach. The basic idea is to involve language information in multilingual training, so that multilingual data can be used in a more effective way. Our study is based on the i-vector architecture and trains a language-aware probabilistic linear discriminative analysis (PLDA) model, which simply treats i-vectors of the same speaker but in different languages as different classes during PLDA training. By the language-aware PLDA, speakers are represented by different latent factors when they speak in different languages, leading to a more discriminative representation. This approach is largely motivated by the phone-aware methods in speaker recognition. For example, Larcher et al. [9] proposed to use phone information when constructing the WCCN [10] matrix. In [11], the authors presented a text-aware PLDA, which treats a single speaker as different classes when he/she speaks different phrases, leading to discriminant on both speakers and phrases. This is similar to the language-aware PLDA presented here, though we focus on discriminating both speakers and languages.

The rest of this paper is organized as follows: Section 2 briefly introduces the i-vector/PLDA framework, and Section 3 proposed the language-aware PLDA training. Section 4 presents the experimental results, and Section 5 concludes the paper.

## 2 I-vector/PLDA framework

### 2.1 I-vector

By the i-vector model, a speaker supervector $M$ is assumed to be a linear Gaussian of the form:

$$M = m + Tw \tag{1}$$

where $m$ is the mean supervector of the UBM, $T$ is the total variability matrix, and $w$ is a low-dimensional vector that represents the whole speech utterance. The prior of $w$ is assumed to be a normal distribution. Given a set of training speech signals $\{X_i\}$, the model training is cast to maximizing the following objective function with respect to the loading matrix $T$:

$$\mathcal{L}(T) = \sum_i ln\{P(X_i;T)\} = \sum_i ln\{\sum_M P(X_i;M)P(M;T)\}$$

where the conditional probability $P(X_i;M)$ is modelled by a GMM, and the prior probability $P(M;T)$ is a Gaussian. Once $T$ is estimated, inferring the posterior probability of $w$ given an utterance $X$ is simple since $P(w|X)$ is a Gaussian as well. In most cases only the mean vector of the posterior is concerned. This is a maximum a posterior (MAP) estimation, and leads to the i-vector of the utterance $X$. More details can be found in [12].

During test, the i-vectors of the enrollment and test utterances inferred by MAP, and the score that the two utterances are from the same speaker can be computed as the cosine distance between the two i-vectors.

### 2.2 PLDA

I-vectors represent both speaker and non-speaker variabilities such as channels and noise. In order to promote the discriminative capability among speakers, various normalization approaches have been proposed, among which PLDA shows clear advantage [13]. Let $w_{(r,n)}$ denotes the i-vector of the $r$-th utterance from $n$-th speaker, the PLDA model forms a linear Gaussian generative process as follows:

$$w_{(r,n)} = m + \mathbf{V}y_n + \mathbf{U}x_{(r,n)} + \epsilon_{(r,n)} \tag{2}$$

where $m$ is a global vector, and $y_n$ and $x_{(r,n)}$ represents the speaker factor and the session factor, respectively, and $\epsilon_{(r,n)}$ represents the residual. The loading matrices $V$ and $U$ define the speaker subspace and the channel subspace, respectively. The factors $y_n$ and $x_{(r,n)}$ are assumed to follow a prior of normal distribution, and $\epsilon_{(r,n)}$ follows a Gaussian distribution whose mean is $\mathbf{0}$ and covariance $\mathbf{\Sigma}$. The

expectation-maximization (EM) algorithm can be used to estimate the model parameters $\{m, \mathbf{U}, \mathbf{V}, \mathbf{\Sigma}\}$, while the MAP estimation can be used to infer the speaker vector $y_n$.

### 2.3 PLDA scoring

Although PLDA scoring can be conducted by inferring the speaker factor $y_n$ and then computing the cosine score, a more systematic way is by hypothesis test. Given two i-vectors $w_1$ and $w_2$, the confidence that they belong the same speaker can be computed as the likelihood ratio as follows:

$$LR = log\frac{P(w_1, w_2|H_{true})}{P(w_1, w_2|H_{imp})} \tag{3}$$

where $H_{true}$ is the hypothesis that $w_1$ and $w_2$ are from the same speaker, and $H_{imp}$ represents the hypothesis that the two i-vectors are from different speakers. This approach marginalizes over $y_n$, and therefore is more accurate.

## 3 Language-aware PLDA training

In multilingual PLDA training, each speaker is regarded as a single class, in spite the language of the training utterances. A consequence of this multilingual training is that the PLDA model can implicity deal with multilingual enrollment and test utterances: the underlying speaker factor can be inferred by considering the possibility that the input utterances are in each of the languages that have been covered by the model. Although this inference does not really happen in the PLDA scoring (Eq. 3), the speaker factor perspective helps understand the strength of the multilingual training. This is shown in Fig. 1, where the circles represent the contour of the distribution of the MAP speaker factor $y_n$. We use solids circles to denote the contours of $y_n$ inferred from utterances in language L1, and dot circles to denote the contours of $y_n$ inferred from utterances in language L2. In plot (a), the PLDA model is trained with data in L1, while in plot (b), it is trained with data in both L1 and L2. It can be found that when inferring speaker factors for utterances in L2 (the dot circles), the L1-trained PLDA leads to a large inter-speaker overlap, due to the uncertainty of the model on the new language. For the multilingual-trained PLDA, the discriminant on both L1 and L2 is clear, even though the enrollment and test utterances are from different languages.

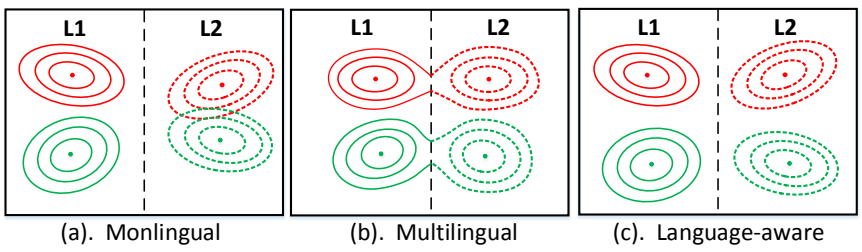


**Figure 1** Distribution of latent speaker factors with PLDA different training methods.

In spite the improved capability in handling multiple language, the discriminative power on a particular language is inevitably decreased by the multilingual training. This is because the distribution of the MAP speaker factor has to be 'broader' (precisely, larger trace of the covariance) to cover multiple languages, leading to 'probability dispersion'. We propose a language-aware training approach to solve this problem. Specifically, the class definition used in the PLDA training involves both the speaker and language identities. In other words, utterances of the same speaker are regarded as in different classes if they are spoken in different languages. By this simple change, the PLDA model learns to discriminate both speaker and language in its latent space, resulting in highly distinct posteriors for the same speaker in different languages. This is shown in plot (c) of Fig 1. It should be emphasized that this language-aware training is different from language-dependent training, where separated PLDA models are trained for each language. The language-aware training is still a multilingual training, so the statistical strength with multilingual data is retained, leading to a stronger model compared to language-dependent models that are trained on data of individual languages. Finally, we note that the language-aware training cannot solve the enrollment-test mismatch. From Fig 1 plot (c), it is clear that i-vectors of utterances from the same speaker but in different languages are clearly separated in the latent speaker factor space, leading to a large intra-speaker variation hence weak inter-speaker discrimination.

## 4 Experiments

### 4.1 Data and settings

The speech database used in this study is CSLT-CUDG2014 [14], a Chinese-Uyghur bi-lingual speech corpus created by CSLT@Tsinghua University. This database involves two languages: Mandrin Chinese and Uyghur, which are used as two official languages in Xinjiang Uyghur autonomous regions of China. This database is designed to study the effect of language mismatch so the discrepancy caused by other factors is intentionally excluded, including linguistic content, channel, noise, emotion, etc. The speech signals were recorded in the sampling rate of 16 kHz and the sample size if 16 bits, using a single smart phone. The contents of the recordings are Chinese and Uyghur digital strings.

There are 181 speakers in the database. For each speaker, two enrollment speech segments, one in Chinese and one in Uyghur, were recorded. For test, each speaker recorded about 10 speech segments in each language for each speaker. Each enrollment segment lasts 40-60 seconds and each test segment lasts 2-3 seconds. The 181 speakers are split into two sets. The first is the training set, which involves 2816 utterances in Chinese or Uyghur from 130 speakers. This set is used to train the $UBM$ model, the $T$ matrix and the PLDA. The rest 51 speakers comprise the evaluation (test) set, used to test the system performance. In our previous work [14], only the female part of the CSLT-CUDG2014 database was used to study the language mismatch challenge. This work used both the female and male data.

The experimental system was based on the i-vector/PLDA framework. The feature was 20-dimensional Mel Frequency Cepstral Coefficients (MFCCs) plus their $\Delta$ and $\Delta\Delta$ derivatives. The utterance-level cepstral mean and variance normalization (CMVN) was employed to remove the channel effect, and an energy based voice

activity detection (VAD) was applied to remove unvoiced segments from the speech. The number of Gaussian components in the UBM was set to 128 and the i-vector dimension was set to 400. The Kaldi toolkit [15] was used to perform the model training and test.

### 4.2 Baseline System

The baseline system is a multilingual system with the conventional PLDA. Specifically, the UBM and the loading matrix of the i-vector model are trained with all the Chinese and Uyghur data. For PLDA, we tested three scenarios: Chinese PLDA and Uyghur PLDA, trained with the Chinese and Uyghur data respectively, and multilingual PLDA, trained using both the Chinese and Uyghur data.

**Table 1** Baseline EER results

| PLDA | Enrollment | Test. | EER (%) |
|---|---|---|---|
| Chinese | Chinese | Chinese | **2.40** |
| Chinese | Chinese | Uyghur | 6.60 |
| Chinese | Uyghur | Uyghur | 2.99 |
| Chinese | Uyghur | Chinese | 8.98 |
| Uyghur | Chinese | Chinese | 3.59 |
| Uyghur | Chinese | Uyghur | 5.59 |
| Uyghur | Uyghur | Uyghur | 3.39 |
| Uyghur | Uyghur | Chinese | 9.98 |
| Multilingual | Chinese | Chinese | 2.60 |
| Multilingual | Chinese | Uyghur | **4.59** |
| Multilingual | Uyghur | Uyghur | **2.99** |
| Multilingual | Uyghur | Chinese | **6.59** |

The results in terms of equal error rate (EER) are presented in Table 1. There are two mono-lingual conditions and two cross-lingual conditions. For each condition, the number of trials is $25,551$, including 501 true speaker trials and $25,050$ imposter trials. The best results in each condition are highlighted. It can be seen that the multilingual trained PLDA exhibits clear advantage compared to the monolingual trained ones, particularly in the cross-lingual test conditions. In the Chinese-Chinese condition, the Chinese PLDA shows a little superior (2.40 v.s. 2.60), while in the Uyghur-Uyghur condition, the multilingual PLDA is better (3.39 v.s. 2.99). This inconsistent results confirms our conjecture that the effect of the multilingual training is two-fold: on one hand, it can utilize more data, and on the other hand, the uncertainty on speakers is increased due to the introduction of other languages.

### 4.3 Language-aware training

The results with the language-aware PLDA training are shown in Table 2, where 'LA' denotes 'language-aware training'. For comparison, the EER results with the conventional multilingual training are also presented. It can be seen that in the two monolingual test conditions, the language-aware training achieves better performance than the conventional multilingual training. For the Chinese-Chinese test and the Uyghur-Uyghur test, the relative EER reduction is 15.38 % and 20.07 %,

respectively. Interestingly, the results with the language-aware PLDA are even better than those obtained with the monolingual PLDAs. This strongly supports our conjecture that the language-aware training can leverage the advantage of multilingual data and avoid the problem of probability dispersion. Finally, we found that the cross-lingual results with the language-aware PLDA are worse than the conventional multilingual PLDA. This again conforms our conjecture that language-aware training leads to large inter-speaker variation hence worse cross-lingual performance.

**Table 2** EER results with language-ware PLDA

| | | EER (%) | |
|---|---|---|---|
| Enrollment | Test. | No LA | LA |
| Chinese | Chinese | 2.60 | **2.20** |
| Chinese | Uyghur | 4.59 | 4.99 |
| Uyghur | Uyghur | 2.99 | **2.39** |
| Uyghur | Chinese | 6.59 | 7.78 |

## 5 Conclusion

This paper presented a language-aware PLDA training approach for multilingual speaker recognition. With the language information involved, the PLDA model can largely avoid the probability dispersion problem while still take the advantage of multilingual training in statistical strength. Our experiment in a Chinese-Uyghur multilingual speaker recognition task showed that the proposed method obtained a relative EER reduction of 15.38% in the Chinese test and 20.07% in the Uyghur test, which validated the idea of language-aware training. Future work will test the method on larger data sets to confirm its strength.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No.61271389 and NO.61371136 and the National Basic Research Program (973 Program) of China under Grant No.2013CB329302.

**Author details**

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

**References**

1. William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 210–229, 2006.
2. Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.
3. Taufiq Hasan and John H. L. Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 4, pp. 842–853, 2013.
4. Bin Ma and Hellen Meng, "English-Chinese bilingual text-independent speaker verification," in *ICASSP 2004*. IEEE, 2004.
5. R. Auckenthaler, M.J. Carey, and J.S.D. Mason, "Language dependency in text-independent speaker verification," in *ICASSP 2001*. IEEE, 2001.
6. Thomas F. Quatieri Douglas A. Reynolds and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
7. Abhinav Misra and John H. L. Hansen, "Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora," in *IEEE Spoken Language Technology Workshop*. IEEE, 2014, pp. 372–377.

8. Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–797, 2011.
9. Kong Aik Lee1 Anthony Larcher, Pierre-Michel Bousquet and Driss Matrouf, "Ivectors in the context of phonetically-constrained short utterances for speaker verification," in *ICASSP 2012*. IEEE, 2012, pp. 4773–4776.
10. Sachin Kajarekar Andrew O. Hatch and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *INTERSPEECH 2006*. IEEE, 2006, pp. 1471–1474.
11. Bin Ma Anthony Larcher, Kong Aik Lee, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *ICASSP 2013*. IEEE, 2013, pp. 7673–7677.
12. Lukáš Burget Ondřj Glembek and Pavel Matějka, "Simplification and optimization of i-vector extraction," in *ICASSP 2011*. IEEE, 2011, pp. 4516–4519.
13. Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *INTERSPEECH 2011*, 2011, pp. 249–252.
14. Rozi Askar, Dong Wang, and Fanhu Bie, "Cross-lingual speaker verification based on linear transform," in *IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, 2015, pp. 519–523.
15. Ghoshal A Povey D and Boulianne G et al, "The kaldi speech recognition toolkit," in *Proc of ASRU*, 2011.