Investigating In-domain Data Requirements for PLDA Training

Md Hafizur Rahman1, David Dean2, Ahilan Kanagasundaram3, Sridha Sridharan4

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{m20.rahman3, a.kanagasundaram3, s.sridharan4}@qut.edu.au, ddean@ieee.org2

Abstract

This paper analyzes the limitations upon the amount of in-domain (NIST SREs) data required for training a probabilistic linear discriminant analysis (PLDA) speaker verification system based on out-domain (Switchboard) total variability subspaces. By limiting the number of speakers, the number of sessions per speaker and the length of active speech per session available in the target domain for PLDA training, we investigated the relative effect of these three parameters on PLDA speaker verification performance in the NIST 2008 and NIST 2010 speaker recognition evaluation datasets. Experimental results indicate that while these parameters depend highly on each other, to beat out-domain PLDA training, more than 10 seconds of active speech should be available for at least 4 sessions/speaker for a minimum of 800 speakers. If further data is available, considerable improvement can be made over solely out-domain PLDA training.

Index Terms: speaker verification, PLDA, in-domain, out-domain, in-domain data requirement

1. Introduction

Over the last few years the state-of-the-art text independent speaker verification has been greatly influenced by cosine similarity scoring (CSS) i-vector and probabilistic linear discriminant analysis (PLDA) [1, 2], which resulted in excellent performance on recent NIST speaker recognition evaluations (SREs). But for any domain other than the standard NIST SREs, current speaker verification systems performance are not satisfactory if sufficient development/training speech data is not available in the target domain.

One of the key requirements to achieve state-of-the-art speaker verification performance is the use of huge amount of speech data during development [3]. But in real world application it is very difficult and often unrealistic to gather huge amount of speech data to develop a state-of-the-art speaker verification system. Most of the previous research focused only on finding effect of limiting the number of speakers for PLDA speaker verification [4, 5]. But in reality not only the number of speakers but also the number of sessions per speaker and the active length of the speech have a direct influence on the development of speaker verification systems.

Recent studies focused on short utterances with different speaker verification techniques like joint factor analysis (JFA) [6], support vector machines (SVM) [7] and PLDA [8, 9] showed that speaker verification performance degrades when short utterances are used for evaluation. Recently, Kanagasundaram et al. [10] studied the speaker verification performance with limited number of sessions, limited number of speakers of microphone data [11] and proposed techniques for reliable estimation of PLDA parameters.

In 2013, the Speaker and Language Recognition Workshop at the Johns Hopkins University (JHU) [12], introduced the Domain Adaptation Challenge, designed to evaluate the system performance built on out-domain data (data not from NIST SREs). Preliminary results presented at that workshop showed that PLDA system trained on the SWB dataset produces higher error rate than a PLDA system trained on earlier NIST datasets. To cope with this challenge of adapting two different datasets in the i-vector domain, methods like inter dataset variability compensation (IDVC) [13, 14], with-in class covariance correction (WCC) [15] have been proposed recently. Also, Garcia-Romero et al. [4] proposed four supervised PLDA domain adaptation techniques. They also proposed agglomerative hierarchical clustering (AHC) [5] for unsupervised PLDA domain adaptation. Villalba et al. [16] proposed Bayesian adaptation of PLDA with limited data. In most of these previous research in-domain data were used to capture domain variation and adapt out-domain PLDA parameters to improve the system performance. However there has not been any detail investigation on in-domain data requirement for PLDA training instead of using huge out-domain data for training.

In this paper, we closely investigated the performance of LDA-projected, length-normalized, Gaussian PLDA (GPLDA) speaker verification systems when trained on limited in-domain training data. We analysed the performance of the system while decreasing the number of speakers, sessions/speaker and active length of sessions of the development data. We tried to determine the minimum in-domain PLDA training data requirement to beat the out-domain baseline PLDA speaker verification performance.

This paper is structured as follows: Section 2 gives a brief overview of limited data development. Section 3 and 4 details i-vector feature extraction techniques, LDA and length normalized GPLDA system correspondingly. The experimental setup and corresponding results are given in Section 5, Section 6 and Section 7 respectively. Finally, Section 8 concludes the paper.

2. Limited Data Development

2.1. Background

Recent advances in speaker verification systems like i-vector based PLDA modelling has shown much promise that resulted in very low error rates. However, we definitely need a huge amount of speech data for faithful speaker verification. But in practical application it is not easy to find such dataset for system development. Also, we may have to use speaker verification system in adverse environmental conditions, of which we may have very small amount of data. So, we can not hope to
produce good verification result in every condition. One of the way to deal with this problem is to investigate the effect of limited development data in the target domain. Limiting development data does not only mean limiting the number of speakers but also sessions per speaker and active length of each session. So, we need to find out the minimum number speaker as well as sessions per speaker and active length of each session required for marginal or faithful verification.

2.2. Reducing Data Requirements

It is obvious that system performance is going to degrade if we limit the amount of speech data available for development. So it is necessary to match the limited in-domain data with data-enriched out-domain data to produce best performance. One of the ways is to capture the in-domain and out-domain data variation in i-vector space and train PLDA parameters with out-domain data [13, 14, 17]. Another way is to adapt out-domain PLDA parameters with in-domain PLDA parameters [4, 5]. But instead of using huge amount of out-domain data for PLDA training, limited amount of in-domain data can be used to perform better than out-domain PLDA speaker verification system. It is necessary to learn the minimum amount of in-domain data required for PLDA training that performs as well as out-domain PLDA system. The basic block diagram of this system is shown in Figure 1.

3. I-vector Feature Extraction

Instead of using two separate subspaces for speaker and channel like JFA, Dehak et al. [18] proposed a single subspace based presentation of GMM super-vector called the i-vector approach. The reason behind this single subspace representation is some speaker discriminant information remain in the channel subspace which could be useful for speaker modelling. A channel and speaker dependent GMM super-vector can be represented as follows,

\[ \mu_s = m + Tw \]

where \(m\) is the speaker and session independent background UBM super-vector, and \(T\) is a low rank matrix representing the directions of variability across all data. \(w\) is the total-variability factors which is normally-distributed \(N(0,1)\). A detail explanation of training \(T\), and i-vector extraction process is described in [19, 18].

4. Linear Discriminant Analysis (LDA)

The LDA transformation finds new spatial axes that minimize the intra-class variance caused by channel effects and maximize the variance between speakers. Which are calculated as follows,

\[ S_b = \sum_{s=1}^{S} n_s (\bar{w}_s - \bar{w})(\bar{w}_s - \bar{w})^T \]

\[ S_w = \sum_{s=1}^{S} \sum_{i=1}^{n_s} (w_i - \bar{w}_s)(w_i - \bar{w}_s)^T \]

where \(S\) is the total number of speakers, \(n_s\) is the number of sessions of speaker \(s\), \(\bar{w}_s\) is the mean i-vector for each speaker and \(\bar{w}\) is the mean of all speakers are defined by,

\[ \bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} w_i \]

\[ \bar{w} = \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{n_s} w_i \]

where \(N\) is the total number of sessions. The transformation matrix \(G\) can be determined through eigenvalue decomposition of \(S_b^{-1} S_w\). Finally, LDA projected i-vector can be calculated as follows,

\[ w_{LDA} = G^T w \]

5. Length-normalized GPLDA System

In this paper we employed the length-normalized GPLDA system, which is introduced by Garcia-Romero et al. [2]. This approach transforms the non-Gaussian i-vector behaviour into Gaussian i-vector behaviour. This technique consists of linear whitening and length normalization. A speaker and channel dependent whitened and length-normalized i-vector can be defined as,

\[ w_{LDA-w} = w_{LDA} + U_1 x_1 + U_2 x_2 + \epsilon_r \]

where for given speaker recordings \(r = 1,2,..,R\), \(w_{LDA} + U_1 x_1\) is the speaker specific part and \(U_2 x_2 + \epsilon_r\) is the channel specific part; The covariance matrix of the speaker part is \(U_1 U_1^T\) and the covariance matrix of the channel part is \(U_2 U_2^T + A^{-1}\).
Scoring in GPLDA is calculated using the batch likelihood ratio \([20]\) between target i-vectors \(w_{LDA\rightarrow target}\) and test i-vectors \(w_{LDA\rightarrow test}\), as follows,
\[
\ln \frac{P(w_{LDA\rightarrow target}, w_{LDA\rightarrow test} | H_1)}{P(w_{LDA\rightarrow target} | H_0)P(w_{LDA\rightarrow test} | H_0)} = (8)
\]
where \(H_1\): The speakers are same, \(H_0\): The speaker are different.

6. Experimental Setup

6.1. Training Datasets

6.1.1. Out-domain Dataset (SWB)

The SWB dataset serves as out-domain dataset throughout our experiments. This dataset consists of standard telephone calls taken from Switchboard-I and Switchboard-II and consists of 1115 male, 1231 female speakers and 22,318 sessions data. Figure 2 shows the cumulative distribution of SWB dataset. In our experiment we used full SWB dataset for out-domain PLDA training.

6.1.2. In-domain Dataset (NIST)

Our in-domain dataset consists of subset of telephone calls taken from NIST 2004, 2005 and 2006 SRE corpora. This dataset consists of 1467 male, 2004 female speakers and 30,317 sessions data. From Figure 2 it can be observed that most of speakers in the dataset have at least 8 sessions and only half of them has more than 8 sessions.

To create the data scarce condition of the PLDA training phase we created several sub-lists from our in-domain dataset. Sub-lists with same amount speakers contain different amount of sessions and sub-lists with same amount of sessions contain different amount speakers. For our experimental purpose we looked at a range of PLDA training speaker counts between 1400 and 100 (per gender) and session counts between 8 and 2 (per speaker).

6.2. System Setup

We extracted 13 dimensional feature-warped MFCCs with appended delta coefficients from raw speech data and used two gender dependent UBMs with 512 Gaussian mixtures in our experiments. The UBMs were trained on telephone speech data from SWB. Baum-Welch statistics were calculated using these UBMs before training a gender-dependent total-variability sub-space.

An i-vector extractor with \(R_{iv} = 500\) is used which is trained on switchboard data. Prior to GPLDA training we reduced i-vector dimension to 150 using LDA projection. We used length-normalized i-vector for GPLDA \([2]\) modelling with i-vector centering and whitening. GPLDA parameters were trained on subsets from NIST 2004, 2005 and 2006 SRE corpora. For GPLDA, best 120 eigenvoices were selected to produce better speaker verification performance. Rather than using full precision matrix, \(A\), we used the diagonal. For score normalization we used source normalisation (S-normalization) \([21]\) in our experiments and normalization dataset was formed by pooling random utterance from NIST 2004, 2005 and 2006 SRE corpora. For evaluation we used NIST 2010 10sec-10sec evaluation condition and the performance is evaluated using the equal error rate (EER).

7. Results and Discussions

Figure 3: Performance comparison of out- and in-domain PLDA speaker verification systems on the common set of the NIST-2010 10sec-10sec evaluation conditions when out-domain PLDA is trained using standard SWB data and in-domain PLDA is trained using different number of sessions/speaker and active speech, performance are presented with respect to active speech length.

Figure 4: Performance comparison of out- and in-domain PLDA speaker verification systems on the common set of the NIST-2010 10sec-10sec evaluation conditions when out-domain PLDA is trained using standard SWB data and in-domain PLDA is trained using different number of speakers and active speech, performance are presented with respect to active speech length.

Figure 3 compares the performance of out-domain and in-domain PLDA speaker verification on the common set of the NIST-2010 10sec-10sec evaluation conditions when out-domain PLDA is trained using standard SWB data and in-domain PLDA is trained using different lengths of active speech.
with limited number of sessions/speaker NIST data. It can be observed with the aid of Figure 3 that at least 4 sessions/speaker are required for in-domain PLDA training to perform better than an out-domain system. It was also observed that in-domain PLDA speaker verification achieves best performance when PLDA is trained using 30 seconds of NIST data. Subsequently, we experimented with the in-domain PLDA approach using different length of active speech and different numbers of speakers in order to find out minimum utterance length. The performance comparison is shown in Figure 4. It can be observed that we require at least 500 speaker with 30 seconds active speech length for PLDA training to estimate reliable PLDA parameters. Based upon results from Figures 3 and 4 we conclude that short length utterance are adequate to train PLDA though PLDA is trained using different number of sessions/speaker or speakers. In real world environments, though it is hard to collect huge amount of in-domain data, our experiments studies have found that short length (30sec) in-domain data is adequate for PLDA training.

Finally, the performance of out-domain PLDA speaker verification is compared against in-domain PLDA approach when in-domain PLDA is trained using 30 sec NIST data with different number of speakers and sessions/speaker. Figure 5(a) illustrates the performance comparison with respect to the number of sessions/speaker and Figure 5(b) illustrates the performance comparison of the same result with respect to number of speakers. It can be clearly observed that system performance goes down with limiting the number of sessions/speaker (Figure 5(a)) and limiting the number of speakers (Figure 5(b)) for in-domain PLDA training. It was also found that at least 800 speakers should be available for 4 sessions/speaker development, at least 500 speakers should be available for 6 sessions/speaker development and at least 500 speakers should be available for 8 sessions/speaker development. It was also observed that when PLDA is trained using 2 sessions/speaker, system achieves the worst performance as at least 4 sessions/speaker data with at least 800 speakers is required for reliable PLDA parameter estimation.

8. Conclusions

In this paper, we presented the results of in-domain data requirement for faithful PLDA training. We performed experiments to find out how the PLDA system performance varies when PLDA is training using short utterance development data, and find out the minimum number of in-domain speakers as well as number of sessions/speaker required to produce acceptable speaker verification performance. We found through experimental studies that 30 second utterances are adequate to train in-domain PLDA approach when evaluated on 10 second utterances. It was also found that for in-domain PLDA training, at least 800 speakers were required for 4 sessions/speaker development, and at least 500 speakers were required for 6 and 8 sessions/speaker development.

9. Acknowledgements

This project was supported by an Australian Research Council (ARC) Linkage grant LP130100110.

10. References


