Investigating In-domain Data Requirements for PLDA Training

Md Hafizur Rahman[1], *David Dean*[2], *Ahilan Kanagasundaram*[3], *Sridha Sridharan*[4]

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{m20.rahman[1], a.kanagasundaram[3], s.sridharan[4]}@qut.edu.au, ddean@ieee.org[2]

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