Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 32–57

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Robust discriminative training against data insufficiency in PLDA-based speaker verification☆

Johan Rohdin*, Sangeeta Biswas, Koichi Shinoda

Tokyo Institute of Technology, Japan

Received 3 December 2014; received in revised form 6 June 2015; accepted 8 June 2015
Available online 20 June 2015

Abstract

Probabilistic linear discriminant analysis (PLDA) with i-vectors as features has become one of the state-of-the-art methods in speaker verification. Discriminative training (DT) has proven to be effective for improving PLDA's performance but suffers more from data insufficiency than generative training (GT). In this paper, we achieve robustness against data insufficiency in DT in two ways. First, we compensate for statistical dependencies in the training data by adjusting the weights of the training trials in order for the training loss to be an accurate estimate of the expected loss. Second, we propose three constrained DT schemes, among which the best was a discriminatively trained transformation of the PLDA score function having four parameters. Experiments on the male telephone part of the NIST SRE 2010 confirmed the effectiveness of our proposed techniques. For various number of training speakers, the combination of weight-adjustment and the constrained DT scheme gave between 7% and 19% relative improvements in $\hat{C}_{\text{llr}}$ over GT followed by score calibration. Compared to another baseline, DT of all the parameters of the PLDA score function, the improvements were larger.

Keywords: Speaker verification; PLDA; Discriminative training; Statistically dependent training data; Overfitting

1. Introduction

In recent years, the combination of i-vector (Dehak et al., 2009, 2011) and probabilistic linear discriminant analysis (PLDA) (Ioffe, 2006; Kenny, 2010) has become one of the state-of-the-art systems in speaker verification. In this system, utterances are mapped into low dimensional vectors known as i-vectors. An i-vector contains information related to the speaker identity as well as irrelevant factors such as speaker's emotions, transmission channels, languages, and environmental noise. Given two i-vectors, the PLDA model separates speaker factors from irrelevant factors and provides a log-likelihood ratio (LLR) score for the two i-vectors being from the same speaker or not.

The PLDA parameters are usually optimized by generative training (GT) under the maximum likelihood (ML) criterion. However, several studies have suggested that discriminative training (DT) is beneficial, either as a complement or as an alternative to GT (Brümmer, 2010; Burget et al., 2011; Cumani et al., 2011, 2012, 2013; Borgström and McCree,

☆ This paper has been recommended for acceptance by Murat Saraclar.

* Corresponding author. Tel.: +81 35734 3481.
E-mail address: johan@ks.cs.titech.ac.jp (J. Rohdin).

http://dx.doi.org/10.1016/j.csl.2015.06.003
0885-2308/

2013). In particular, score calibration by means of a discriminatively trained affine transformation (AT-Cal) (Brümmer, 2010), has become popular. AT-Cal only adjusts the scores and can therefore be applied to any speaker verification system. DT schemes that are specific to PLDA have also been proposed. A DT scheme that optimizes all the parameters of the PLDA LLR score function (Scr-UC)[1] was proposed by Burget et al. (2011) and Cumani et al. (2011) and a DT scheme that optimizes the PLDA model parameters instead of its score function, was proposed by Borgström and McCree (2013). However, DT is in general less robust against data insufficiency than GT. For example, in Cumani and Laface (2014), Scr-UC was worse than GT when the number of training speakers was less than around 1600. In this paper, we tackle the data insufficiency problem in two approaches. One is to effectively use the limited amount of training data. The other is to constrain the model parameters to avoid overfitting.

When the amount of training data is limited, each training utterance or speaker is often used in more than one training trial in the model training. Accordingly, the training trials are not statistically independent. As a consequence, the *average loss* of the training trials that we use as training objective is not the best estimate of the *expected loss*. We propose to adjust the weights of the training trials in order to obtain the *best linear unbiased estimator* (BLUE) of the *expected loss*.

In order to find the constraints that best avoid overfitting without constraining the model too much, we propose three discriminative training schemes that are less constrained than Src-UC (Burget et al., 2011; Cumani et al., 2011) but more flexible than AT-Cal (Brümmer, 2010). The first is a transformation of the PLDA LLR score function having four parameters. The second is a scaling of each element in the i-vectors. The third is a training scheme that, like Src-UC, updates all parameters of the PLDA LLR score function but preserves some properties of PLDA that are removed by Scr-UC (Rohdin et al., 2014a). Experiments on the male telephone part of the NIST SRE 2010 confirmed the effectiveness of our proposed techniques.

The remainder of this paper is organized as follows. Section 2 introduces the necessary background including the detection cost function, i-vector and PLDA based speaker-verification and discriminative PLDA training. Section 3 performs an analysis of the discriminative training methods. Based on the conclusions in Section 3, Section 4 presents the compensation for the statistical dependence, and Section 5 presents constrained discriminative PLDA training. Section 6 experimentally evaluates the methods. Finally, Section 7 concludes this paper.

2. Background

2.1. Detection cost function

When making a decision based on the score from a speaker verification system, it is typically desired to minimize the expected cost of the decision. This is reflected in the *detection cost function* (DCF) used in the NIST evaluations. When the test and enrollment utterances in a trial are from the same speaker, we refer to the trial as a *target trial*, otherwise we refer to it as a *non-target trial*. The DCF measures the cost for an application with a prior probability of a target trial, P_{tar}, and the costs C_{FR} and C_{FA} for false rejection (FR) and false acceptance (FA) respectively.

$$\text{DCF} = P_{\text{tar}} C_{\text{FR}} P_{\text{FR}} + (1 - P_{\text{tar}}) C_{\text{FA}} P_{\text{FA}}, \tag{1}$$

where $P_{\text{FR}} = P(\text{error} \mid \text{target})$ and $P_{\text{FA}} = P(\text{error} \mid \text{non} - \text{target})$ are the empirical probabilities for FR and FA respectively estimated in the evaluation database. For the purpose of ranking systems, a scaling of the DCF does not make any difference. Therefore, for system optimization it is equivalent to use

$$\text{DCF}' = P_{\text{eff}} P_{\text{FR}} + (1 - P_{\text{eff}}) P_{\text{FA}}, \tag{2}$$

where

$$P_{\text{eff}} = \frac{P_{\text{tar}} C_{\text{FR}}}{P_{\text{tar}} C_{\text{FR}} + (1 - P_{\text{tar}}) C_{\text{FA}}}, \tag{3}$$

[1] UC refers to *unconstrained*.