Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language xxx (2018) xxx-xxx

www.elsevier.com/locate/csl

Nonparametrically trained PLDA for short duration *i*-vector speaker verification☆

Abbas Khosravani, Mohammad M. Homayounpour*

Laboratory for Intelligent Multimedia Processing, Department of Computer Engineering and Information Technology, Amirkabir University of Technology, 424 Hafez St, Tehran, Iran

Received 10 June 2017; received in revised form 29 October 2017; accepted 28 December 2017
Available online xxx

Abstract

The duration of speech segments can significantly impact the performance of text-independent speaker verification systems. In real world applications which require high accuracy on short utterances, the performance of *i*-vector speaker verification framework degrades significantly considering that *i*-vectors extracted from short utterances are less reliable (i.e., uncertainty is higher) than those extracted from long utterances. Therefore, to handle duration variability properly, a more realistic approach seems to be required. This study is an extension to our recently proposed nearest neighbor probabilistic linear discriminant analysis (NN-PLDA) which estimates the parameters of PLDA in *i*-vector speaker verification framework using a nonparametric form rather than maximum likelihood estimation (MLE) obtained by an EM algorithm, and has been shown to provide superior performance. In NN-PLDA, the between-speaker covariance matrix that represents global information about the speaker variability is replaced with a local estimation computed on a nearest neighbor basis for each target speaker. Compared to their parametric counterparts, the nonparametric between- and within-speaker scatter matrices can better exploit the discriminant information in training data and are more adapted to sample distributions. In this paper, we provide further analysis on the proposed nonparametrically trained PLDA as well as introduce a duration variability modeling technique in the estimation of the within-speaker scatter matrix as to compensate for the effect of limited speech data. We evaluate our approach using *core−10sec* and *10sec−10sec* telephone trial conditions of NIST 2010 SRE as well as on the truncated test utterances in extended core condition with duration less than 10 s. We also present the results obtained by the successful incorporation of NN-PLDA on the recent NIST 2016 speaker recognition evaluation. In all experiments, considerable performance improvement is obtained with the proposed technique compared to a generatively trained PLDA model.

Keywords: Speaker recognition; PLDA; Nonparametric; NIST SRE; Short duration; *i*-Vector

1. Introduction

Despite the fact that highly reliable speaker recognition performance can be obtained when sufficient amounts of speech are available for enrolment and/or verification, securing satisfactory performance with limited amount of

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author.
E-mail address: homayoun@aut.ac.ir (M.M. Homayounpour).

http://dx.doi.org/10.1016/j.csl.2017.12.009
0885-2308/ 2017 Published by Elsevier Ltd.

speech which is required in many real world applications, has become the focus of speaker recognition community. The significant degradation in *i*-vector speaker verification performance with short duration utterances for enrolment and/or verification is due to the fact that *i*-vectors extracted from short utterances are less reliable (i.e., more uncertainty) than those extracted from long utterances as they vary due to speaker, session and linguistic content (Kenny et al., 2013). To address this issue, a considerable number of research efforts have studied techniques for the compensation of the duration variability in *i*-vector space or through score calibration (Hasan et al., 2013; Mandasari et al., 2011; Sarkar et al., 2012; Kanagasundaram et al., 2011, 2016). Experimental studies have found that partitioning long enrolled utterances into multiple short utterances and averaging, improves probabilistic linear discriminant analysis (PLDA) speaker verification (Kanagasundaram et al., 2016). In Kanagasundaram et al. (2014) authors proposed a short utterance variance normalization (SUVN) technique to compensate for the mismatch between short utterances and their longer counterparts. This technique works by capturing duration variability in a short utterance variance (SUV) matrix through truncating long utterances in a large set of development data. The SUV matrix is then used to compensate for the variation between short- and full-length utterances by transforming the *i*-vectors into a SUV-projected space. In Hasan et al. (2013), it has also been shown that duration variability can be considered as additive noise in the *i*-vector space and based on this assumption, a multi-duration training by PLDA has been found to improve performance. In Kenny et al. (2013) it has been shown that propagating the uncertainty associated with the *i*-vector extraction process into a PLDA classifier could lead to substantial improvements in accuracy under variable length test utterances, however, the likelihood ratio computation for speaker verification requires posterior distributions of the *i*-vectors, making it computationally more expensive than the standard PLDA and would probably be too unwieldy to experiment with.

We recently proposed a nonparametric technique to estimate the parameters of the PLDA in *i*-vector speaker verification framework as an alternative to maximum likelihood estimation (MLE) obtained by an EM algorithm and showed its superior performance on long duration utterances (Khosravani and Homayounpour, 2017a). The proposed nearest neighbor PLDA (NN-PLDA) technique is inspired by the recent success of the nonparametric discriminant analysis (NDA) (Fukunaga and Mantock, 1983) in speaker recognition (Sadjadi et al., 2014, 2016). In the proposed approach the between-speaker covariance matrix that represents global information about the speaker variability is replaced with a local estimation for each target speaker and is obtained using speakers with the most similarity to that target speaker. The new formulation in which both the within-speaker and between-speaker scatter matrices are redefined in nonparametric form, can better exploit the discriminant information in training data. Moreover, they lead to a model more adapted to sample distribution especially in non-Gaussian case of *i*-vectors without length-normalization. The estimated parameters will be used to compute verification score in the same way as in generative PLDA model. An estimation of the verification score using a discriminative model rather than a generative model has been proposed in Burget et al. (2011). In this approach the speaker verification score for a pair of *i*-vectors is computed using the same functional form derived from the PLDA generative model but the parameters are estimated using a discriminative training criterion that discriminates between same-speaker and different-speaker trials rather than using maximum likelihood estimation (MLE) of PLDA model parameters (Brummer, 2010). Training can be performed by means of support vector machines (SVMs) and a suitable kernel derived from the PLDA generative model (Cumani et al., 2011). However, it has been shown that discriminative training of probabilistic models needs more training data and only provides competitive performance with the ones obtained by generative models (Rohdin et al., 2014; Cumani et al., 2013).

In this study, we aim at providing further analysis on the proposed NN-PLDA model as well as introducing a duration variability modeling technique in the estimation of within-speaker scatter matrix so as to compensate for the effect of limited speech data. Our main contributions are to show that using a simple parametric form of the scatter matrices, we can achieve almost the same performance as of G-PLDA. We then show that using a nonparametric form of between-speaker scatter matrix which explicitly emphasizes the speakers near boundary contributes to the improvement of speaker verification performance.

The remainder of this paper is organized as follows. Section 2 details the *i*-vector extraction mechanisms and processing. Section 3 outlines a typical state-of-the-art Gaussian PLDA modeling technique for speaker verification. Our proposed NN-PLDA technique with duration variability modeling technique is presented in Section 4. In Section 5, we describe experimental setup on the protocols defined by NIST with results and analysis in Section 6. Finally, Section 7 concludes the paper.