Improving Short Utterance based I-vector Speaker Recognition using Source and Utterance-Duration Normalization Techniques

A. Kanagasundaram[⋆]*, D. Dean*[⋆]*, J. Gonzalez-Dominguez*[†]*,*
S. Sridharan[⋆]*, D. Ramos*[†]*, J. Gonzalez-Rodriguez*[†]

[⋆] Speech Research Laboratory, Queensland University of Technology, Australia
[†] ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
[⋆]{a.kanagasundaram, d.dean, s.sridharan}@qut.edu.au,
[†] {javier.gonzalez, daniel.ramos, joaquin.gonzalez} @uam.es

Abstract

A significant amount of speech is typically required for speaker verification system development and evaluation, especially in the presence of large intersession variability. This paper introduces a source and utterance-duration normalized linear discriminant analysis (SUN-LDA) approaches to compensate session variability in short-utterance i-vector speaker verification systems. Two variations of SUN-LDA are proposed where normalization techniques are used to capture source variation from both short and full-length development i-vectors, one based upon pooling (SUN-LDA-pooled) and the other on concatenation (SUN-LDA-concat) across the duration and source-dependent session variation. Both the SUN-LDA-pooled and SUN-LDA-concat techniques are shown to provide improvement over traditional LDA on NIST 08 truncated 10sec-10sec evaluation conditions, with the highest improvement obtained with the SUN-LDA-concat technique achieving a relative improvement of 8% in EER for mis-matched conditions and over 3% for matched conditions over traditional LDA approaches.

Index Terms: speaker verification, i-vector, total-variability, LDA, WCCN

1. Introduction

In the presence of intersession variability conditions, significant amount of speech is required for speaker verification system development and current state-of-the-art systems still require substantial amounts of speech for training and testing. Reducing the amount of speech required for development, training and testing while obtaining satisfactory performance has been the focus of a number of recent studies in state-of-the-art speaker verification design, including joint factor analysis (JFA), i-vectors, probabilistic linear discriminant analysis (PLDA) and support vector machines (SVM) [1, 2, 3, 4]. Recently, Kenny *et al.* [5], have investigated how to quantify the uncertainty associated with the i-vector extraction process and propagate it into a PLDA classifier.

As an i-vector approach is based on defining only one variability space [6, 7], instead of the separate session variability and speaker spaces of the JFA approach, it is believed that i-vectors will not lose significant speaker information in session variability space [7], and is also believed that this would be additional advantage to short utterance speaker verification. Until now, several inter-session variability compensation approaches have been introduced to a long utterance cosine similarity scoring (CSS) i-vector speaker recognition system [7, 8, 9]. However, no single inter-session variability compensation approach has been specially analyzed with short utterance ($\sim$10s) based CSS i-vector speaker recognition. As our main focus of this paper is to analyze the inter-session variability compensation approaches with short utterances, we have chosen the simple CSS over PLDA classifier, as CSS is computationally more efficient approach than PLDA.

In this paper, initially, we analyze how the short utterance based i-vector system performs when the standard session variability compensation approaches, including LDA and WCCN are trained using long- and short-length utterance development data. Later, we also introduce source and utterance-duration normalized LDA (SUN-LDA) approach for the purpose of improving speaker verification performance under short utterance evaluation conditions.

This paper is structured as follows. Section 2 gives a brief introduction to speaker verification using i-vectors. Section 3 details the proposed SUN-LDA approach. The experimental protocol and corresponding results are given in Section 4 and Section 5.

2. Speaker verification using i-vectors

In contrast to the separate speaker and session dependent subspaces of JFA, i-vectors represent the GMM super-vector using a single total-variability subspace [6]. An i-vector speaker and session dependent GMM super-vector can be represented by

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{w}, \quad (1)$$

where $\mathbf{m}$ is the speaker and session independent background UBM super-vector, $\mathbf{T}$ is a low rank matrix representing the primary variation across a large collection of development data. $\mathbf{w}$ is normally distributed with parameters $N(\boldsymbol{0}, \boldsymbol{I})$, and is the *i-vector* representation used for speaker verification. R_w is dimension of i-vector features.

The total-variability subspace, $\mathbf{T}$, training and subsequent i-vector extraction is detailed by Dehak *et al.* [10]. As we are investigating the telephone and microphone based speaker verification system in this paper, the total-variability subspace should be trained in a manner which exploits the useful speaker variability contained in speech acquired from both telephone and microphone sources. In this paper, the total-variability subspace ($R_w^{\text{telmic}} = 500$) is trained on telephone and microphone speech pooled utterances, as recent studies have found that pooled total-variability approach is better than concatenated total-variability approach [11].