Short Utterance Variance Modelling and Utterance Partitioning for PLDA Speaker Verification

Ahilan Kanagasundaram, David Dean, Sridha Sridharan, Clinton Fookes, Ivan Himawan

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{a.kanagasundaram, d.dean, s.sridharan, c.fookes, i.himawan}@qut.edu.au

Abstract

This paper analyses the short utterance probabilistic linear discriminant analysis (PLDA) speaker verification with utterance partitioning and short utterance variance (SUV) modelling approaches. Experimental studies have found that instead of using single long-utterance as enrolment data, if long enrolled-utterance is partitioned into multiple short utterances and average of short utterance i-vectors is used as enrolled data, that improves the Gaussian PLDA (GPLDA) speaker verification. This is because short utterance i-vectors have speaker, session and utterance variations, and utterance-partitioning approach compensates the utterance variation. Subsequently, SUV-PLDA is also studied with utterance partitioning approach, and utterance-partitioning-based SUV-GPLDA system shows relative improvement of 9% and 16% in EER for NIST 2008 and NIST 2010 truncated 10sec-10sec evaluation condition as utterance-partitioning approach compensates the utterance variation and SUV modelling approach compensates the mismatch between full-length development data and short-length evaluation data.
Index Terms: speaker verification, i-vectors, PLDA, SUV, utterance partitioning

1. Introduction

A significant amount of speech is required for speaker model enrolment and verification, especially in the presence of large intersession variability, which has limited the widespread use of speaker verification technology in everyday applications. 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]. Continuous research on this field has been ongoing to address the robustness of speaker verification technologies under such conditions.

Previous research studies had found that long utterance i-vectors contain two source of variation: changing speaker characteristics, and changing channel (or session) characteristics [5]. Recently it was found that short utterance i-vectors vary due to speaker, session and linguistic content (utterance variation) [6]. In typical PLDA speaker verification, a single utterance is used as enrolment data. In this paper, instead of using a single long-utterance as enrolment data, long-enrolment utterances are partitioned into multiple short-enrolment utterances and the average i-vector over the short utterances is used as enrolment data to improve the short utterance speaker verification system. Recently, we have also introduced short utterance variance (SUV) modelling to PLDA speaker verification system to compensate the session and utterance variations [7]. Subsequently, in this paper, we also investigate the utterance-partitioning-based Gaussian PLDA (GPLDA) speaker verification with SUV modelling approach.

This paper is structured as follows: Section 2 details the i-vector feature extraction techniques. Section 3 details the short utterance variance added i-vector feature extraction approach. Section 4 explains the GPLDA based speaker verification system. The experimental protocol and corresponding results are given in Section 6 and Section 7. Section 8 concludes the paper.

2. I-vector feature extraction

I-vectors represent the GMM super-vector by a single total-variability subspace. This single-subspace approach was motivated by the discovery that the channel space of JFA contains information that can be used to distinguish between speakers [8]. An i-vector speaker and channel dependent GMM super-vector can be represented by,

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{m}$ is the same universal background model (UBM) super-vector used in the JFA approach and $\mathbf{T}$ is a low rank total-variability matrix. The total-variability factors ($\mathbf{w}$) are the i-vectors, and are normally distributed with parameters *N(0,1)*. Extracting an i-vector from the total-variability subspace is essentially a *maximum a-posteriori adaptation* (MAP) of $\mathbf{w}$ in the subspace defined by $\mathbf{T}$. An efficient procedure for the optimization of the total-variability subspace $\mathbf{T}$ and subsequent extraction of i-vectors is described Dehak *et al.* [5, 9]. In this paper, the pooled total-variability approach is used for i-vector feature extraction where the total-variability subspace (dimension = 500) is trained on telephone and microphone speech utterances together.

3. Short utterance variance

The long-length utterance i-vectors have speaker and session variations whereas short-length i-vectors have speaker, session and a lot of utterance variations. Thus, during development for SUV-PLDA, utterance variance is captured using the inner product of the difference between the full- and short-length i-vectors, and it is artificially added to full-length utterances and the simulated SUV is modelled using the PLDA approach [10]. The short utterance variance matrix, $\mathbf{S}_{SUV}$, can be calculated as follows,

$$\mathbf{S}_{SUV} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{A}^T(\mathbf{w}_n^{full} - \mathbf{w}_n^{short}))(\mathbf{A}^T(\mathbf{w}_n^{full} - \mathbf{w}_n^{short}))^T \tag{2}$$