i-vector Based Speaker Recognition on Short Utterances

Ahilan Kanagasundaram, Robbie Vogt, David Dean, Sridha Sridharan, Michael Mason

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{a.kanagasundaram, r.vogt, d.dean, s.sridharan, m.mason}@qut.edu.au

Abstract

Robust speaker verification on short utterances remains a key consideration when deploying automatic speaker recognition, as many real world applications often have access to only limited duration speech data. This paper explores how the recent technologies focused around total variability modeling behave when training and testing utterance lengths are reduced. Results are presented which provide a comparison of Joint Factor Analysis (JFA) and i-vector based systems including various compensation techniques; Within-Class Covariance Normalization (WCCN), LDA, Scatter Difference Nuisance Attribute Projection (SDNAP) and Gaussian Probabilistic Linear Discriminant Analysis (GPLDA). Speaker verification performance for utterances with as little as 2 sec of data taken from the NIST Speaker Recognition Evaluations are presented to provide a clearer picture of the current performance characteristics of these techniques in short utterance conditions.

Index Terms: speaker verification, short utterance, i-vector, SDNAP, WCCN, LDA, Gaussian PLDA

1. Introduction

The significant amount of speech required for speaker model enrolment and verification, especially in the presence of large intersession variability, has limited the widespread use of speaker verification technology in everyday applications. Continuous research on this field has been ongoing to address the robustness of speaker verification technologies under such conditions. Reducing the amount of speech required while obtaining the satisfactory performance has been the focus in a number of recent studies focused on Joint Factor Analysis (JFA) and SVM based speaker verification. These studies have shown that while performance degrades considerably in very short utterances ($< 10s$) for both approaches, JFA [1] appears to a better choice in these conditions than SVMs [2]. This paper will focus on whether a recently proposed factor-analysis front-end approach to speaker verification, called *i-vectors* [3], could form a suitable foundation for continuing research into short utterance speaker verification.

JFA originally proposed by Kenny [4], has recently evolved as a powerful tool in speaker verification to model the inter-speaker variability and to compensate for channel/session variability in the context of high-dimensional Gaussian Mixture Model (GMM) supervectors. More recently a new front-end factor analysis technique, termed i-vector (for intermediate-size vector) extraction, proposed by Dehak *et al.*[5], has evolved from JFA. Rather than taking the JFA approach of modelling a speaker and channel variability space, the i-vector approach forms a low-dimensional total-variability space that models both speaker and channel variability. The i-vector approach proposed in [3] also has the advantage that scoring uses a simple Cosine Similarity Scoring (CSS) kernel directly to perform verification, making the scoring process faster and less complex than other speaker verification methods, including JFA or Support Vector Machines (SVM) supervector approaches. Because the total variability space does contain channel variability information, i-vector speaker verification systems need to be combined with intersession compensation techniques such as Within-class Covariance Normalisation (WCCN), Linear Discriminant Analysis (LDA) and Nuisance Attribute Projection (NAP) [3]. More recently, Kenny *et al.* [6] have developed a new technique called Gaussian Probabilistic LDA (GPLDA), which divides the i-vector space into speaker and session variability subspaces, which has shown significant promise for intersession compensation for i-vector speaker verification.

The main aim of this paper is to investigate the effect that short duration utterances have on both enrolment and training when using the i-vector approach. As this approach is based on defining only one variability space, instead of the separate channel and speaker spaces of the JFA approach, we expect that i-vectors won't loose any speaker information with reduction of utterance duration [7]. We also report on the short utterance duration performance when various intersession variability compensation techniques such as WCCN, LDA and NAP are used in conjunction with i-vectors, including a short investigation of scatter-difference NAP (SDNAP), which has not yet been considered for i-vector compensation. In addition we compare the above combination of techniques with GPLDA [6]. The experimental results presented give useful indication of how the systems degrade as training and testing utterance lengths are reduced.

Section 2 of the paper provides a condensed introduction to each of the techniques investigated, from JFA, through the various front end factor analysis techniques used with total variability modeling and finally GPLDA. Section 3 provides details of the experimental procedures used to investigate the short utterance performance and presents the outcomes of the three experiments for discussion.

2. Factor analysis speaker verification

Factor analysis approaches to speaker verification were originally intended to model the intersession variability directly in the construction of the super-vectors used for scoring verification trials, such as in the standard JFA approach [4]. More recently factor analysis techniques have been considered as a front-end to form a low-dimensional total-variability subspace that can be classified more efficiently than the high-dimensional supervectors used in JFA and SVM speaker verification systems. This section will briefly outline the JFA approach, followed by a more detailed description of the common approaches

28 – 31 August 2011, Florence, Italy