SPEAKERS IN THE WILD (SITW): The QUT Speaker Recognition System

H. Ghaemmaghami, M. H. Rahman, I. Himawan, D. Dean, A. Kanagasundaram,
S. Sridharan and C. Fookes

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia
{houman.ghaemmaghami, m20.rahman, i.himawan}@qut.edu.au, ddean@ieee.org,
{a.kanagasundaram, s.sridharan, c.fookes}@qut.edu.au

Abstract

This paper presents the QUT speaker recognition system, as a competing system in the *Speakers In The Wild* (SITW) speaker recognition challenge. Our proposed system achieved an overall ranking of second place, in the main *core-core* condition evaluations of the SITW challenge. This system uses an i-vector/PLDA approach, with domain adaptation and a deep neural network (DNN) trained to provide feature statistics. The statistics are accumulated by using class posteriors from the DNN, in place of GMM component posteriors in a typical GMM-UBM i-vector/PLDA system. Once the statistics have been collected, the i-vector computation is carried out as in a GMM-UBM based system. We apply domain adaptation to the extracted i-vectors to ensure robustness against dataset variability, PLDA modelling is used to capture speaker and session variability in the i-vector space, and the processed i-vectors are compared using the batch likelihood ratio. The final scores are calibrated to obtain the calibrated likelihood scores, which are then used to carry out speaker recognition and evaluate the performance of the system. Finally, we explore the practical application of our system to the *core-multi* condition recordings of the SITW data and propose a technique for speaker recognition in recordings with multiple speakers.

1. Introduction

Recent advances in speaker recognition technology, specifically text-independent speaker verification, have brought about significant gains in system accuracy. The joint factor analysis (JFA) speaker modelling approach proposed by Kenny [1], has evolved into a powerful tool for speaker verification. This is because the JFA approach allows for modelling of inter-speaker variability and compensation for channel/session variability in the context of high-dimensional Gaussian mixture model (GMM) supervectors. This technique advanced to a new front-end factor analysis technique, termed i-vector (for intermediate-size vector) extraction, proposed by Dehak *et al.* [2]. In this technique, rather than taking the JFA approach of modelling a speaker and channel variability spaces, a low-dimensional total-variability space that models both speaker and channel variability is trained. The i-vector approach proposed in [2], has the advantage of scoring using a 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).

The use of i-vectors for speaker modelling has been established as the state-of-the-art approach to representing speech segments produced by a speaker identity, however i-vectors are susceptible to unwanted variations due to mismatch of linguistic content and recording channel information between segments of speech spoken by the same speaker identity. To overcome this, a range of techniques, such as within-class covariance normalisation (WCCN), linear discriminant analysis (LDA) and nuisance attribute projection (NAP) were proposed and shown to be effective for this purpose [3]. Kenny [4] then introduced the use of the PLDA approach for modelling channel variability within the i-vector space. The PLDA technique was originally proposed by Prince *et al.* [5] for face recognition, and later it was adapted for modelling the i-vector distributions for speaker verification [4]. Two PLDA approaches, Gaussian PLDA (GPLDA) and heavy-tailed PLDA (HTPLDA) were introduced [4, 6]. Garcia-Romero *et al.* [6], demonstrated that the heavy-tailed behaviour of i-vector features can be converted into Gaussian behaviour by using a length-normalized approach, and thus length-normalized GPLDA was demonstrated to achieve similar performance to the HTPLDA technique.

Until recently the use of a GMM universal background models (UBM), trained on large amounts of unlabelled data for capturing phonetic variations of speech in an unsupervised manner, was the state-of-the-art approach and an essential part of speaker recognition technology. The successful application of deep neural network (DNN) acoustic models to the task of automatic speech recognition (ASR) [7] and the significant gains in accuracy that resulted from the use of a DNN for this task, has brought about proposals for the use of such a DNN in the task of speaker recognition [8, 9]. In order to apply such an ASR DNN to the task of speaker recognition, the GMM-UBM is replaced by a DNN to collect sufficient statistics for i-vector speaker modelling [9]. The input to the DNN is frequency-domain features, such as the mel-frequency cepstral coefficient (MFCC) features, while the output of the DNN provides soft alignments for phonetic content, in the form of tied triphone states, referred to as senones [8]. This technique has resulted in significant improvements over GMM-UBM i-vector systems [8]. It is hypothesised that this is due to the ability of the DNN to model phonetic content/variations directly, as opposed to the unsupervised expectation-maximization (EM) approach in the GMM-UBM training process for capturing some acoustic patterns in the data, which would highly depend on the training data. This however is achieved at the cost of higher computational complexity and can put significant strain on resources.

Another factor that can significantly impact the performance of speaker verification systems is the mismatch between audio domains of the training and test data. The performance variation due to cross-domain speaker verification, was first addressed at the Summer Workshop at Johns Hopkins University (JHU) held in 2013 [10]. Results presented in that