A Study of X-vector Based Speaker Recognition on Short Utterances

A. Kanagasundaram[1,2]*, S. Sridharan*[2]*, G. Sriram*[3]*, S. Prachi*[3]*, C. Fookes*[2]

[1]Department of Electrical & Electronic Engineering, University of Jaffna, Sri Lanka
[2]Speech and Audio Research Lab, SAIVT, Queensland University of Technology, Australia
[3]LEAP Lab, Indian Institute of Science, India

ahilan@eng.jfn.ac.lk, sriramg@iisc.ac.in, c.fookes,s.sridharan@qut.edu.au

Abstract

The aim of this work is to gain insights into how the deep neural network (DNN) models should be trained for short utterance evaluation conditions in an x-vector based speaker verification system. The study suggests that the speaker embedding can be extracted with reduced dimensions for short utterance evaluation conditions. When the speaker embedding is extracted from deeper layer which has lower dimension, the x-vector system achieves 14% relative improvement over baseline approach on EER on NIST2010 5sec-5sec truncated conditions. We surmise that since short utterances have less phonetic information speaker discriminative x-vectors can be extracted from a deeper layer of the DNN which captures less phonetic information. Another interesting finding is that the x-vector system achieves 5% relative improvement on NIST2010 5sec-5sec evaluation condition when the back-end PLDA is trained using short utterance development data. The results confirms the intuitive expectation that duration of development utterances and the duration of evaluation utterances should be matched. Finally, for the duration mismatch condition, we propose a variance normalization approach for PLDA training that provides a 4% relative improvement on EER over baseline approach.

Index Terms: Speaker verification, PLDA, DNN, x-vector, speaker embedding, short utterance

1. Introduction

The development of state-of-the-art speaker verification systems for short utterances is an active area of research since potential users of the system prefer short utterance for enrollment and authentication. For wider development of speaker verification technology, high accuracy need to be achieved with short utterances; however current state-of-the-art such as x-vector based systems still require significant amount of speech for enrollment and verification. Researchers have been working to improve the performance of speaker verification on short utterance evaluation conditions [1, 2, 3, 4].

Previously, joint factor analysis (JFA), i-vector, probabilistic linear discriminant analysis (PLDA) based speaker recognition systems were studied on short utterances [1, 5, 2, 3, 4]. These studies have shown that when the evaluation utterance length is reduced, it significantly affects the performance [1, 2, 4].

Recently, the deep learning approaches have been incorporated into i-vector-based speaker recognition systems using two main approaches: (1) A speech-based Deep Neural Network (DNN) is used to extract bottleneck (BN) features from the middle layer restricted in dimensionality [6]; (2) DNN senone approach, where the calculation of Baum-Welch statistics is based on speech-based DNN [7, 8]. Though DNN senone based speaker verification systems achieve the state-of-the-art performance, this approach is computationally expensive and it is infeasible to use in practical applications. More recently, researchers proposed the end-to-end x-vector speaker recognition systems [9, 10, 11]. In the end-to-end x-vector approach, deep neural network is fed with a variable length utterance and maps it to a speaker embedding [10]. Snyder *et al* proposed data augmentation to achieve further improvement on DNN speaker embedding based speaker recognition [12]. Existing DNN architectures work well with long utterances but the performance drops when the utterance length is reduced [12].

In this paper, we investigate approaches for improving the performance of time delay DNN architecture in x-vector based speaker verification under short utterance evaluation conditions. Since short utterances are likely to contain less phonetic information compared to long utterances, we hypothesize that it is sufficient to use lower dimensional speaker embedding which can capture all the variations present in the short utterances. In an x-vector based speaker recognition system, the speaker embedding is normally extracted from the layer adjacent to the stats pooling layer of the DNN as this high dimensional layer is expected capture all the speaker discriminative information at phonetic level. Since the phonetic information in short utterances is small, we argue that it would be sufficient to capture speaker embedding from deeper low dimensional layers of the DNN.

We also investigate the training of the PLDA which is used as the back-end of the x-vector based speaker verification system. When PLDA is trained using full-length data and the speaker recognition system is evaluated on short-length data, the data duration mismatch significantly affects the performance. To overcome this mismatch, the PLDA can be trained using short-length data. However, this causes mismatch when evaluated on longer utterances. In this paper, novel utterance variance transformation is proposed to compensate the data duration mismatch and improve the performance of under utterance length mismatch evaluation conditions.

This paper is structured as follows: Section 2.1 details the extraction of speaker embedding features. Section 2.2 provides the details of PLDA classifier. The proposed short utterance variance transformation is detailed in Section 2.3. The experimental protocol and corresponding results are given in Section 3 and Section 4. Section 5 concludes the paper.