A Study on Pairwise LDA for X-vector based Speaker Recognition

A. Kanagasundaram, S. Sridharan, S. Ganapathy and C. Fookes

In typical x-vector based speaker recognition systems, standard linear discriminant analysis (LDA) is used to transform the x-vector space with the aim of maximizing the between-speaker discriminant information while minimizing the within-speaker variability. For LDA, it is customary to use all the available speakers in the speaker recognition development dataset. In this study, we investigate if it would be more beneficial to estimate the between-speaker discriminant information and the within-speaker variability using the most confusing samples and the most distant samples (from the target speaker mean) respectively in the LDA based channel compensation. The between-speaker variance is estimated using a pairwise approach where the most confusing non-target speaker samples are found based on the Euclidean distance between the speaker mean and adjacent speaker's samples. The within-speaker variance is estimated using the mean of each speaker and the furthermost samples in the speaker sessions. Experimental results demonstrate the proposed LDA approach for an x-vector x-vector based speaker recognition system achieves over 17% relative improvement on EER over standard LDA based x-vector speaker recognition systems on the NIST2010 corext-corext condition.

Introduction

In recent times, research in speaker recognition has focused on deep learning based approaches. Deep learning approaches have been incorporated into i-vector-based speaker recognition systems using two main approaches: (1) A speech-based DNN is used to extract bottleneck (BN) features from the middle layer [1]. Instead of mel frequency cepstral coefficient (MFCC) features, the BN features are used in i-vector speaker recognition systems; and (2) DNN senone approach, where the calculation of Baum-Welch statistics is based on a speech-based DNN [2]. Even though DNN senone based speaker verification systems achieves state-of-the-art performance, the approach is computationally more expensive and its use is limited in practical applications [3].

More recently, the end-to-end speaker recognition systems and x-vector speaker recognition systems have become popular in speaker recognition research [4, 5]. In the x-vector/ speaker embedding based speaker recognition systems, the variable length speaker utterance is mapped to fixed-length x-vectors using a deep neural network [4]. Once the x-vectors are extracted using a deep neural network, the standard linear discriminant analysis (LDA) approach is used to transform the x-vector. The LDA compensates the channel mismatch by maximizing the between-speaker discriminant information (between-speaker variance) while minimizing the within-speaker variability (within-speaker variance). Finally, the length normalized Gaussian probabilistic linear discriminant analysis (GPLDA) is used as a backend to calculate the scoring [6].

There has been a few recent efforts to improve the performance of LDA for speaker verification, however, these have been based mainly on i-vector based speaker verification systems. These efforts have included: source-normalized LDA [7] which normalize the scatter matrices across different sources; the weighted LDA [8] which weights class pairs in an inverse proportion to their distance; nonparametric discriminant analysis (NDA) [9] which calculates both within- and between-class scatter matrices on a local basis using a nearest neighbour rule; and local pairwise LDA [10] which maximizes the local pairwise covariance, which represents the local structure between the target class samples and neighboring non-target class samples.

In this work, we focus on improving LDA based channel compensation in x-vector based speaker recognition. Specifically, we investigate whether it would be more beneficial to estimate between-speaker discriminant information in LDA using the most confusing samples and within speaker variability using the most distant samples as opposed to estimating these variances using all the available speaker samples in the development set as is routinely performed. Our investigation reveals that the proposed approach can provide a significant reduction in error rate compared to the standard LDA in x-vector based speaker verification.

This paper is structured as follows; Section 1 provides a brief introduction to x-vector based speaker recognition systems. Section 2 initially details the standard LDA based channel compensation approach and also subsequently introduces the proposed channel compensation approach. The experimental protocol and corresponding results are given in Section 3 and Section 4. Section 5 concludes the paper.

1. X-vector based speaker recognition systems

A feed-forward DNN is used to compute the x-vectors from speech samples of variable utterance length. Once fixed length x-vectors are extracted from speech segments, the LDA based channel compensation is used to increase the between-speaker variability and reduce the within-speaker variability. The details of standard and proposed channel compensation approaches are given in Section 2. Finally, a GPLDA based back-end classifier is used to classify the speakers.

1.1. Extraction of x-vector features

The extraction of x-vectors using a feed-forward DNN network is shown in Figure 1. The feed-forward DNN is trained using a large amount of training data to classify the speakers [11]. The first four layers of the networks operate at the frame-level. If the t is the current time step, t-2, t-1, t, t+1, t+2 frames are spliced together at the input layer. As the 23 MFCC features are extracted for our experiments, the input layer dimension is 115. The size of the output layer is 512. In the first hidden layer, t-2, t, t+2 frames are spliced together and the size input is 1536 (512x3). In the second hidden layer, t-3, t, t+3 frames are spliced together and the size of the input is 1536 (512x3). The dimensions of the third and fourth layers are respectively 512 and 1500. The fifth layer is stats pooling where all the frames are aggregated together and the mean and standard deviation are estimated and concatenated together (1500x2). From this layer onward, the utterance level parameters are estimated. The sizes of the sixth and seventh layers have a dimension of 512. Finally, the softmax layer is trained to classify the speakers from the training dataset. The DNN architecture is trained to classify the 3413 speakers in the training dataset. After the network training, the x-vectors (512) are extracted from layer 6.

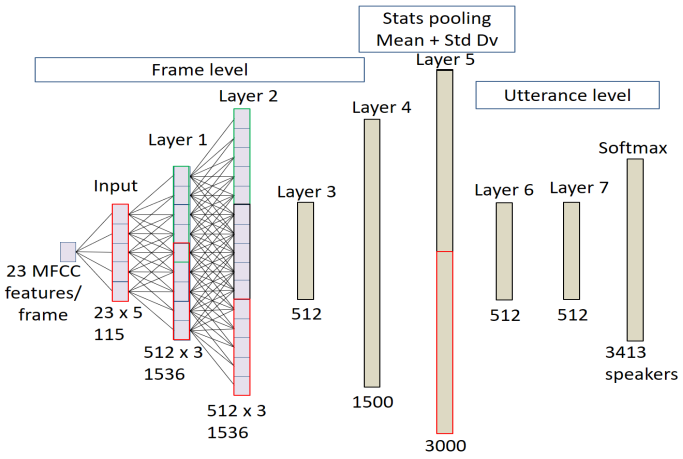

Fig. 1 *A block diagram of the x-vector extractor is shown where the first four layers of network operate at the frame level, the fifth layer is stats pooling, and the sixth and seventh layers of the network operate at the utterance level.*

1.2. PLDA classifier

The length-normalized GPLDA is used as a backend classifier as it is a simplified and computationally efficient approach compared to heavy-tailed PLDA. The length-normalization approach is applied on the development and evaluation data prior to GPLDA modeling [12].

A speaker and channel dependent length-normalized i-vector, $\mathbf{w}_r$, can be defined as,

$$\mathbf{w}_r \quad = \quad \mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \boldsymbol{\varepsilon}_r, \tag{1}$$

where for the given speaker recordings $r = 1, ...R$; $\mathbf{U}_1$ is the eigenvoice matrix and $\mathbf{x}_1$ is the speaker factor and $\boldsymbol{\varepsilon}_r$ is the residual. The covariance matrix, $\mathbf{U}_1\mathbf{U}_1{}^T$, represents the between-speaker variability, and the covariance matrix, $\boldsymbol{\Lambda}^{-1}$, describes the within-speaker variability. The model parameters, $\mathbf{U}_1$, and $\boldsymbol{\Lambda}$ are iteratively estimated using the expectation maximization algorithm. Scoring is calculated using the batch likelihood ratio between a target and test x-vectors.