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.

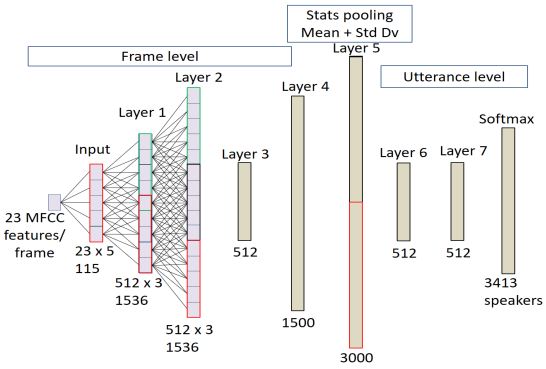

Fig. 1 *A block diagram of the x-vector extractor is shown where the first four layers of network operate at the frame level, the fifth layer is stats pooling, and the sixth and seventh layers of the network operate at the utterance level.*

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.

2. Channel compensation approaches

As the x-vector features have both speaker and channel information, the channel compensation approaches are used to compensate for the within-speaker variability and increase the between-speaker discriminating information. In typical x-vector based speaker recognition systems, standard LDA approaches are used to increase the between-speaker while reducing the within-speaker variability [6]. In standard LDA, all the speakers and all their sessions that are available in the development data set are used to estimate these two variances.

2.1. Standard LDA approach

The between-speaker variance, $\mathbf{S}_b$, and within-speaker variance, $\mathbf{S}_w$, can be calculated as follows,

$$\mathbf{S}_b = \sum_{s=1}^{S} n_s(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})^T, \tag{2}$$

$$\mathbf{S}_w = \sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T, \tag{3}$$

where the mean x-vector for across-all-speakers, $\bar{\mathbf{w}}$, is defined by $\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\mathbf{w}_i^s$. N is the total number of sessions. n_s is the number of available sessions per speaker s and S is the total number of speakers in the development set.

LDA attempts to find a reduced set of dimensions that minimize the within-class variability while maximizing the between-class variability through the eigenvalue decomposition of,

$$\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}. \tag{4}$$

2.2. Proposed pairwise approach

In the standard between-speaker variance estimation, the variance is estimated between each speaker mean $\bar{\mathbf{w}}_s$ and the global mean $\bar{\mathbf{w}}$.

In contrast, the pairwise between-speaker variance is estimated as follows,

$$\mathbf{S}_b^p = \frac{1}{N}\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} n_i n_j(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)^T. \tag{5}$$

In the pairwise between-speaker variance, the variance is estimated between the mean of each pair of speakers i and j. It can be proved that,

$$\mathbf{S}_b^p = \mathbf{S}_b. \tag{6}$$

Therefore, the pair-wise estimation of between-speaker variance estimation depicted in Eq (5) is equivalent to the standard representation in Eq(2). In the rest of manuscript, we will use the pairwise estimation represented by Eq (5) for between-speaker variance.

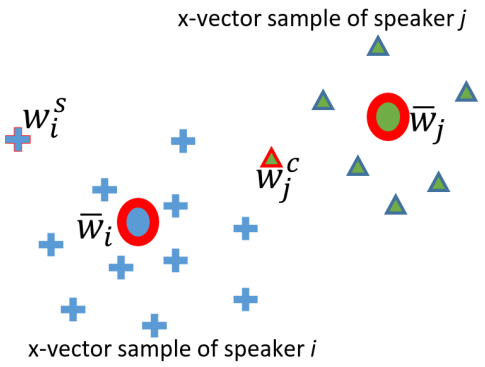

Fig. 2 *The illustration of the proposed pairwise variance estimations. The between-speaker variance is estimated between $\bar{\mathbf{w}}_i$ and $\mathbf{w}_j^c$ where $\mathbf{w}_j^c$ is the non target speaker's sample closest to target speaker mean $\bar{\mathbf{w}}_i$. The within-speaker variance is estimated between $\bar{\mathbf{w}}_i$ and $\mathbf{w}_i^s$ where $\mathbf{w}_i^s$ is the speaker sample furthermost from the target speaker mean $\bar{\mathbf{w}}_i$.*

Our aim is to find out if it would be advantageous to estimate the pairwise between-speaker variance based on the speaker i and the closest sample of speaker j instead of using all the samples of speaker j as customarily computed. The illustration of the proposed pairwise between-speaker variance estimation is shown in Figure 2, and is estimated as follows, where we replace the x-vector mean of the speaker j in Eq (5) with the x-vector value of speaker j's sample which is closest to speaker i,

$$\mathbf{S}_b^p = \frac{1}{N}\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} n_i(\bar{\mathbf{w}}_i - \mathbf{w}_j^c)(\bar{\mathbf{w}}_i - \mathbf{w}_j^c)^T, \tag{7}$$

where $\mathbf{w}_j^k$ is considered as the k^{th} in-class sample of speaker j. The closest sample of the jth speaker is determined as follows: the Euclidean distance between $\bar{\mathbf{w}}_i$ and all the j^{th} speaker's samples $\mathbf{w}_j^k$ are estimated. All the samples are sorted in ascending order based on the Euclidean distance and the closest sample of speaker j to speaker i is selected as $\mathbf{w}_j^c$.

Note that in the pairwise between-speaker variance estimation, the variance is estimated between speaker i and all the pairwise speakers. Some of speakers are closer to speaker i and some others would be far away from speaker i. We wish to determine if it would be advantageous to use only speakers who are close to speaker i in estimating the between-speaker variance. The proposed approach is represented by Eq (8),

$$\mathbf{S}_b^p = \frac{1}{M}\sum_{i=1}^{S-1}\sum_{j=i+1}^{M} n_i(\bar{\mathbf{w}}_i - \mathbf{w}_j^c)(\bar{\mathbf{w}}_i - \mathbf{w}_j^c)^T. \tag{8}$$

The Euclidean distance between $\mathbf{w}_i$ and $\mathbf{w}_j^c$ are estimated and the individual pairwise variance between speaker i and j is sorted in ascending order and the specific percentage of speakers j who are close to i is selected for between-speaker variance estimation. The value of M calculated based on the percentage value. The percentage is given by $(M/S) \times 100$. We will discuss later the choice of the percentages of selected speakers for computing the between-speaker variance.

2.3. Modified within-speaker variance estimation

In the standard within-speaker variance estimation Eq (3), the intra-speaker variance is estimated between each in-class samples, $\mathbf{w}_i^s$ and the mean of speaker s, $\bar{\mathbf{w}}_s$. The within-speaker scatter depends on several factors arising from session to session variations including variations due to microphones, acoustic environments and transmission channels. The objective of the within-speaker variance is to reduce the intra-speaker variability.

We want to determine if it would be advantageous to estimate the between-speaker variance based on the mean of speaker s, $\bar{\mathbf{w}}_s$, and the furthermost samples, $\mathbf{w}_i^s$ instead of using all the samples of speaker s as is performed customarily.

The with-in speaker variance is estimated as follows,

$$\mathbf{S}_w = \sum_{s=1}^{S}\sum_{i=1}^{n_m}(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T. \tag{9}$$

The Euclidean distance between $\mathbf{w}_i^s$ and $\bar{\mathbf{w}}_s$ is estimated. Following this, the individual variances are sorted in descending order based on the Euclidean distance and the specific percentage of the furthermost samples are used for within-speaker variance estimation. The value of n_m is selected based on the percentage value. The percentage is given by $(n_m/n_s) \times 100$ where n_m is the number of speaker samples selected for each speaker s. Again, we will discuss later, the choice of the percentages of samples for computing the intra-speaker variance.

3. Experimental methodology

The experiments were conducted using the kaldi toolkit [13]. The x-vector based experiments were evaluated using the NIST 2010 corpora [14]. For NIST 2010, the performance was evaluated using the equal error rate (EER).

The DNN architecture was trained using NIST and Switchboard data. The NIST data consists of NIST SRE 04, 05, 06, 08, and Switchboard data consists of Switchboard 2 Phases 1, 2, 3 and Switchboard Cellular. The DNN speaker embedding system is trained with 23 MFCCs without their first derivative. The DNN speaker embedding system is trained using the Kaldi recipe. The dimension of x-vector is 512 and the dimension of standard and the proposed LDA approaches is 200. The PLDA classifiers are trained on NIST SRE 04, 05, 06 and 08.

4. Results and discussions

In the initial experiments, the pairwise-between speaker variance is estimated using the most confusing non-target speaker sample for each non-target speaker using the Eq (7). Table 1 compares the EER performance on NIST2010 corext-corext conditions.

Table 1: Performance comparison of x-vector speaker recognition systems on NIST 2010 corext-corext when pairwise between-speaker variance is estimated using the most confusing sample of each non target speaker. Baseline performance is calculated using standard LDA where all the samples are used for the variance estimation as depicted in Eq (2).

Baseline EER	**EER with most confusing sample of all non target speakers**
2.53%	**2.40%**

It can be observed that the system achieves better performance when pairwise between-speaker variance is estimated between the speaker mean of speaker i and the most confusing sample $\mathbf{w}_j^c$.

In Table 2, we provide the results for estimating the pairwise between-speaker variance using speaker i and the most confusing sample of several confusing non-target speakers (M speakers) as depicted by Eq (8). Table 2 compares the EER performance on NIST2010 corext-corext conditions. Note that in this experiment we use the most confusing sample of each of the selected confusing speakers. We estimate the error rate for 3 different percentages.

Table 2: Performance comparison of x-vector speaker recognition systems on NIST 2010 corext-corext when pairwise between-speaker variance is estimated with different percentages of the confusing speakers. The percentage is given by $(M/S) \times 100$ where S is total number of speakers and M is the selected number of confusing speakers.

Baseline	**EER for different amounts of confusing speakers**		
	50%	25%	15%
2.53%	2.27%	2.25%	**2.22%**

It can be observed that the x-vector system achieves better performance when the pairwise between-speaker variance is estimated using 15% of the most confusing samples. The results suggest that it would be better to estimate the pairwise between-speaker variance based on a selected number of confusing non-target speakers who are closer to the target speaker. Note that in Table (2) we couldn't reduce the error rate further by reducing the percentage of selected confusing speakers below 15% since the actual number of confusing speakers available in the database becomes insufficient to provide a stable estimation of the variance of LDA. The results suggest that if larger databases were available with more non-target speakers close to the target speaker, the EER may be further reduced.

In the final experiment, we investigate whether the use of the furthermost samples to estimate the within-speaker variance can provide additional improvements. Table 3 compares the EER performance on NIST2010 corext-corext conditions when the within-speaker variance estimated using Eq (9).

Table 3: Performance comparison of x-vector speaker recognition systems on corext-corext when within-speaker variance estimated using furthermost samples as depicted by Eq (9). The percentage is given by $(n_m/n_s) \times 100$ where n_s is the total number of samples for speaker s and n_m is the selected number of distant samples for speaker s. Note that the between-speaker variance for the LDA is estimated using the best percentage ie:15% of the confusing speakers (based on results in Table (2))

Baseline	**EER for different amounts of furthermost samples**	
	50%	25%
2.53%	2.12%	**2.09%**

It can be observed from the results the x-vector systems achieves better performance when the within-speaker variance is estimated using 25% of the least confusing samples. As noted for Table (2), in Table (3) we couldn't reduce the error rate further by reducing the percentage of selected furthermost samples below 25% since the number of samples available in the database becomes insufficient to provide a stable estimation of the variance for LDA. The results suggest that if larger development databases are available with large number of samples further from the mean for each speaker we may be able to reduce the EER further.

The objective of the within-speaker variance is to reduce the intra-speaker variability. The results suggest that the in-class samples which are close to the speaker mean, do not help to estimate the robust within-speaker variance. The samples which are far away from the speaker mean are the best samples for within-speaker variance estimation.

Through these experimental results we have demonstrated that the proposed LDA for x-vector systems achieves over 17% relative improvement on EER over baseline systems on NIST2010 corext-corext condition. This improvement in EER has been achieved using an intuitively appealing approach, where the pairwise between-speaker variance is estimated using the most confusing sample of the most confusing set of speakers and the within-speaker variance is estimated using the speaker samples that are further away from the speaker mean for each speaker.

5. Conclusion

The results presented in this paper confirms that all speakers are not required to estimate the between-speaker variance and the most confusing samples a play significant role when training the LDA transform. It is also found that the samples which are far away from the speaker mean are the best samples to use for within-speaker variance estimation. Overall the proposed approach achieves a 17% relative improvement on EER over baseline systems on NIST2010 corext-corext condition. This is achieved when the pairwise between-speaker variance is estimated using the most confusing sample of the most confusing speakers and the within-speaker variance is estimated using the samples further away from the speaker mean when implementing the channel compensation for x-vector based speaker recognition using LDA.

A. Kanagasundaram et al (*Speech and Audio Research Lab, SAIVT, Queensland University of Technology, Brisbane, Australia.*)

E-mail: a.kanagasundaram@qut.edu.au

References

1 Richardson, F., Reynolds, D. and Dehak, N.: 'Deep neural network approaches to speaker and language recognition', *IEEE Signal Processing Letters, IEEE*, 2015, 22, pp. 1671-1675.
2 Snyder, D., Garcia-Romero, D. and Povey, D.: 'Time delay deep neural network-based universal background models for speaker recognition', *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 92-97.
3 Garcia-Romero, D. and McCree, A.: 'Insights into deep neural networks for speaker recognition', *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
4 Snyder, D., Garcia-Romero, D., Povey, D. and Khudanpur, S.: 'Deep neural network embeddings for text-independent speaker verification',*Proc. Interspeech*, 2017, pp. 999-1003.
5 Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A. and Zhu, Z.: 'Deep speaker: an end-to-end neural speaker embedding system', *arXiv preprint arXiv*:1705.02304, 2017.
6 Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S.: 'X-vectors: Robust DNN embeddings for speaker recognition', *ICASSP*, 2018.
7 McLaren, M. and van Leeuwen, D.: 'Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources', *Audio, Speech, and Language Processing, IEEE Transactions on, IEEE*, 2012, 20, pp. 755-766.
8 Kanagasundaram, A., Dean, D., Vogt, R., McLaren, M., Sridharan, S. and Mason, M.: 'Weighted LDA Techniques for i-vector Based Speaker Verification', *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 4781-4784.
9 Sadjadi, S. O., Pelecanos, J. and Zhu, W.: 'Nearest neighbor discriminant analysis for robust speaker recognition', *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
10 He, L., Chen, X., Xu, C., Liu, J. and Johnson: M. T. 'Local Pairwise Linear Discriminant Analysis for Speaker Verification', *IEEE Signal Processing Letters, IEEE*, 2018, 25, pp. 1575-1579.
11 Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y. and Khudanpur, S.: 'Deep neural network-based speaker embeddings for end-to-end speaker verification', *Spoken Language Technology Workshop (SLT), 2016 IEEE*, 2016, pp. 165-170.
12 Garcia-Romero, D. and Espy-Wilson, C.: 'Analysis of i-vector length normalization in speaker recognition systems', *International Conference on Speech Communication and Technology*, 2011, pp. 249-252.
13 Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. and others: 'The Kaldi speech recognition toolkit', *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.
14 Martin, A. F. and Greenberg, C. S.: 'The NIST 2010 speaker recognition evaluation', *Eleventh Annual Conference of the International Speech Communication Association*, 2010.