WEIGHTED LDA TECHNIQUES FOR I-VECTOR BASED SPEAKER VERIFICATION

A. Kanagasundaram[1], *D. Dean*[1], *R. Vogt*[1], *M. McLaren*[2], *S. Sridharan*[1], *M. Mason*[1]

[1] Speech Research Laboratory, Queensland University of Technology, Australia
[2] Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[1]{a.kanagasundaram, r.vogt, d.dean, s.sridharan, m.mason}@qut.edu.au, [2]m.mclaren@let.ru.nl

ABSTRACT

This paper introduces the Weighted Linear Discriminant Analysis (WLDA) technique, based upon the weighted pairwise Fisher criterion, for the purposes of improving i-vector speaker verification in the presence of high inter-session variability. By taking advantage of the speaker discriminative information that is available in the distances between pairs of speakers clustered in the development i-vector space, the WLDA technique is shown to provide an improvement in speaker verification performance over traditional Linear Discriminant Analysis (LDA) approaches. A similar approach is also taken to extend the recently developed Source Normalised LDA (SNLDA) into Weighted SNLDA (WSNLDA) which, similarly, shows an improvement in speaker verification performance in both matched and mismatched enrolment/verification conditions. Based upon the results presented within this paper using the NIST 2008 Speaker Recognition Evaluation dataset, we believe that both WLDA and WSNLDA are viable as replacement techniques to improve the performance of LDA and SNLDA-based i-vector speaker verification.

Index Terms— speaker verification, i-vector, linear discriminant analysis

1. INTRODUCTION

Recent research in speaker verification has focused on the i-vector front-end factor analysis technique. This technique was first proposed by Dehak *et al.* [1] to provide an intermediate speaker representation between the high dimensional Gaussian Mixture Model (GMM) supervector and traditional low dimensional feature representations. The extraction of these intermediate-sized vectors, or i-vectors, were motivated by the existing supervector-based Joint Factor Analysis (JFA) approach, but rather than modelling the speaker and channel variability space separately, i-vectors are formed by modelling a single low-dimensional total-variability space that covers both speaker and channel variability [2]. Because channel variability is not explicitly removed in the i-vector extraction approach, channel compensation techniques must be implemented to limit the effects of channel variability in the i-vector speaker representations.

While the choice of channel compensation techniques is very much an active area of research, the use of Linear Discriminant Analysis (LDA) followed by Within Class Covariance Normalization (WCCN) used by Dehak *et al.* [2] has shown good performance. More recently, this approach was extended by McLaren and van Leeuwen [3] by proposing a new LDA-based approach, Source-Normalized LDA (SNLDA), which improve the i-vector-based speaker recognition in both mismatched conditions and conditions for which limited system development speech resources are available.

In this paper, we propose to investigate a new LDA technique, based upon the weighted pair-wise Fisher criteria [4], that has recently shown promise in the field of template-based face recognition [5]. This technique, known as Weighted LDA (WLDA), takes advantage of the discriminatory information between pairs of classes, or speakers for our application, in the between-class scatter that has not yet been investigated for i-vector-based speaker verification. By applying a weighted parameter to class pairs that weights closer pairs higher, WLDA should provide an improvement in discriminative ability between classes that would otherwise be difficult to distinguish in the LDA- or SNLDA-transformed i-vector space. Motivated by the improvements obtained for WLDA over traditional LDA for face recognition [5], our aim in this paper is to investigate if a similar approach can be taken with WLDA and Weighted SNLDA (WSNLDA) to provide improvements for i-vector-based speaker verification.

This paper is structured as follows. Section 2 gives a brief introduction to process of i-vector based speaker verification system. Section 3 details the proposed W-LDA and W-SNLDA techniques. The experimental protocol and corresponding results are given in Section 4 and Section 5.

2. SPEAKER VERIFICATION USING I-VECTORS

In contrast to the separate speaker and channel dependent subspaces of JFA, i-vectors represent the GMM supervector using a single total-variability subspace. This single-subspace approach was motivated by the discovery that the channel space of JFA contains information that can be used to distinguish between speakers [6]. An i-vector speaker and channel dependent GMM supervector can be represented by

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{m}$ is the speaker and channel independent background UBM supervector, $\mathbf{T}$ is a low rank matrix representing the primary directions variation across a large collection of development data. Finally, $\mathbf{w}$ is normally distributed with parameters $N(0, 1)$, and is the *i-vector* representation used for speaker verification.

For details of the total variability subspace training and subsequent i-vector extraction, the reader is encouraged to investigate the techniques covered by Dehak *et al.* [2].

This project was supported by the Cooperative Research Centre for Advanced Automotive Technologies (AutoCRC).

2.1. Channel compensation techniques

As i-vectors are defined by single variability space, containing both speaker and channel information, there is a requirement that additional intersession, or channel variability, compensation approaches be taken before verification. These channel compensation techniques are designed to maximize the effect of between-class variability and minimize the effects of within-class variability due to differences in microphones, acoustic environment and variation in speaker's voices.

While the choice of channel compensation techniques for i-vector representations is very much an active area of research, the use of a LDA-based technique followed by WCCN has been shown to provide a good level of performance [2, 3]. Within this section we will outline the existing LDA + WCCN technique of Dehak *et al.* [2] and the extension into the SNLDA + WCCN technique of McLaren *et al.* [3].

2.1.1. LDA followed by WCCN (LDA + WCCN)

In the first stage of the LDA + WCCN sequential approach, LDA is used to define a new spatial axes $\boldsymbol{A}$ that minimizes the within-class variance caused by channel effects and maximizes the variance between speakers in the i-vector space. WCCN is then used as an additional channel compensation technique to scale the subspace in order to attenuate dimensions of high within-class variance.

Both LDA and WCCN calculations are based up the standard within- and between-class scatter estimations S_w and S_b, calculated as

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

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

where S is the total number of speakers, n_s is number of utterances of speaker s, and N is the total number of sessions. The mean i-vectors, $\bar{\mathbf{w}}_s$ for each speaker, and $\bar{\mathbf{w}}$ is the across all speakers are defined by

$$\bar{\mathbf{w}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{w}_i^s, \quad (4)$$

$$\bar{\mathbf{w}} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\mathbf{w}_i^s. \quad (5)$$

In the first stage, LDA attempts to find a reduced set of axes $\mathbf{A}$ that minimizes the within-class variability while maximizing the between-class variability through the eigenvalue decomposition of $\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}$.

In the second stage, the WCCN transformation matrix ($\boldsymbol{B}$) is trained using the LDA-projected i-vectors [1] from the first stage. The WCCN matrix ($\boldsymbol{B}$) is calculated using Cholesky decomposition of $\boldsymbol{BB}^T = \boldsymbol{W}^{-1}$, where the within-class covariance matrix $\boldsymbol{W}$ is calculated using

$$\mathbf{W} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{A}^T(\mathbf{w}_i^s - \bar{\mathbf{w}}_s))(\mathbf{A}^T(\mathbf{w}_i^s - \bar{\mathbf{w}}_s))^T. \quad (6)$$

2.1.2. SNLDA followed by WCCN (SNLDA + WCCN)

In [3], McLaren*et al.* found that the between-class scatter calculated using the standard LDA approach can be influenced by source variation under mismatched conditions. This influence can be reduced by estimating the between-class scatter using source-normalized i-vectors and fixing the within-class scatter as the residual variations in the i-vector space [3]. The source-normalized between-class scatter, $\mathbf{S}_b^{src}$, can be composed of the source-dependent between-class scatter matrices for telephone and microphone-recorded speech, which can be calculated as follows,

$$\mathbf{S}_b^{src} = \mathbf{S}_b^{tel} + \mathbf{S}_b^{mic} \quad (7)$$

Rather than estimate the within-class scatter separately as in (3), McLaren *et al.* calculated the within-class scatter matrix as the difference between a total variance matrix, $\mathbf{S}_t$, and the source-normalized between-class scatter:

$$\mathbf{S}_w = \mathbf{S}_t - \mathbf{S}_b^{src}, \quad (8)$$

where

$$\mathbf{S}_t = \sum_{n=1}^{N} w_n w_n^t. \quad (9)$$

This approach allows S_w to be more accurately estimated when development dataset do not provide examples of each speech source from every speaker. Similarly to the LDA + WCCN approach outlined previously, after the i-vectors are first projected into the reduced dimensionality SNLDA space, a WCCN matrix is calculated to scale the dimensions in order to minimize the within class covariance.

2.2. Cosine similarity scoring

Scoring of channel-compensated i-vectors for speaker verification is accomplished using a Cosine Similarity Scorer (CSS), which was found to provide similar performance to Support Vector Machine (SVM) based approaches with a considerable increase in efficiency [6]. The CSS operates by comparing the angles between a channel compensated test i-vector, $\hat{w}_{\text{test}}$, and a channel-compensated target i-vector $\hat{w}_{\text{target}}$:

$$\text{score}(\hat{w}_{\text{target}}, \hat{w}_{\text{test}}) = \frac{\langle \hat{w}_{\text{target}}, \hat{w}_{\text{test}} \rangle}{\|\hat{w}_{\text{target}}\| \, \|\hat{w}_{\text{test}}\|}. \quad (10)$$

3. WEIGHTED LDA AND SNLDA

The within and between-class scatter matrices estimated for LDA and SNLDA in the previous section attempt to project high dimensional i-vectors into a more discriminative lower-dimensional subspace. However, these approaches do not take advantage of the discriminative relationships between pairs of classes. This is particularly the case when pairs are positioned closely together, often due to channel similarities, and traditional estimation of between-class scatter matrix are not able to adequately compensate. Intuitively one would surmise that the classes that are closer to each other should be weighted more heavily to reduce class confusion within adjacent class pairs. By applying a pair-wise weighting based upon the pair-wise Fisher criterion, Weighted LDA (WLDA) has been shown to improve template based face recognition [5].

In this section, we will outline how the WLDA technique will be applied to extend both the LDA + WCCN and SNLDA + WCCN i-vector channel compensation approaches outlined in the previous section.

3.1. WLDA followed by WCCN (WLDA + WCCN)

In the WLDA approach, the between-class scatter matrix is redefined by adding a weighting function, $w(d_{ij})$, according to the between-class distance of each pair of classes i and j. This weighted between-class scatter matrix, is defined as

$$\mathbf{S}_b^w = \frac{1}{N}\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} w(d_{ij}) n_i n_j (\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)^T, \qquad (11)$$

where $\bar{\mathbf{w}}_x$, and n_x is the mean i-vector and session count respectively of speaker x.

In (11), the weighting function $w(d_{ij})$ is defined such that the classes that are closer to each other will be more heavily weighted. As we show in Appendix A, when $w(d_{ij})$ equals to 1, the weighted between-class scatter estimations will converge to the standard non-weighted between-class scatter from (2). For this paper, we will investigate two weighting functions, one based on the Euclidean distance, and a second based on the Bayes Error.

The Euclidean distance based monotonically-decreasing weighting function $w_E(d_{ij})$, can be defined as $(d_{ij})^{-n}$ where d_{ij} is the Euclidean distance between the means of i-vector classes i and j, or $\|\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j\|$. The degree parameter n was chosen as 6 for the speaker verification results reported in this paper, based up a limited set of development experiments.

The second weighting parameter was based upon the Bayes Error approximates of the mean accuracy amongst class pairs. The Bayes Error based weighting function $w_B(d_{ij})$), can be calculated as

$$w_B(d_{ij}) = \frac{1}{2(\triangle_{ij})^2}\text{Erf}\,(\frac{\triangle_{ij}}{2\sqrt{2}}), \qquad (12)$$

where $\triangle_{ij}$ is the Mahalanobis distance between the means of classes i and j:

$$\triangle_{ij} = \sqrt{(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)^T(\mathbf{S}_w)^{-1}(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_j)}. \qquad (13)$$

Once the weighted between-class scatter, $\mathbf{S}_b^w$, is estimated for the chosen weighting function, the standard within-class scatter S_w and the corresponding WLDA and WCCN transformation matrices can be estimated and applied as described in the previous section.

3.2. WSNLDA followed by WCCN (WSNLDA + WCCN)

In order to apply the weighting parameter to the SNLDA approach, the source-dependent between-class scatter matrices where calculated using the weighted between-class scatter calculation from (11) and combined to form the source-normalized between-class scatter matrix in the same manner as (7).

However, while in the original SNLDA algorithm, the within-class scatter matrix was estimated as the difference between total variance and the source-normalized between-class variance, this approach is not take for WSNLDA. Because the weighting parameters destroy the relationship between the total variance and the between-class variance, the within-class variance is estimated independently using (3) as in the traditional LDA approach.

4. METHODOLOGY

The proposed methods were evaluated using the NIST 2008 SRE telephone and interview based utterances from the short2-short3 enrol-verification partitions. Performance was evaluated using the equal error rate (EER) and minimum decision cost function (DCF) calculated using $C_{miss} = 10$, $C_{FA} = 1$, and $P_{target} = 0.01$. Evaluation was performed on the NIST 08 DET conditions 3, 4, 5 and 7, corresponding to *interview-interview*, *interview-telephone*, *telephone-interview*, and *telephone-telephone* (English-only) enrolment-verification trials. ZT normalization was applied to all of the experiments with the normalization development data pooled over microphone and telephone sources. Gender pooled results are reported throughout.

Gender-dependent Universal Background Models (UBM), consisting of 512 components were trained on 26-dimensional, feature-warped MFCCs (including deltas) on data taken from the NIST 2004, 2005, and 2006 SRE corpora. These gender-dependent UBMS were used to calculate the Baum-Welch statistics for calculation of a total variability subspace of dimension $R_w = 500$ is to calculate the i-vector speaker representations. The development data for the total variability and channel compensation subspaces, were obtained from NIST 2004, 2005 and 2006 SRE corpora as well as Switchboard II. Both telephone and microphone data was pooled for development. 150 eigenvectors were selected as best value for LDA training by performance on a development dataset.

5. RESULTS AND DISCUSSION

The results of the weighted LDA and SNLDA techniques are shown in comparison to their non-weighted baselines in Tables 1 and 2 respectively. It can be seen that the weighted techniques have generally provided a useful improvement over the non-weighted techniques in all but the *telephone-telephone* conditions, generally regardless of the weighting function chosen. The choice of weighting function appears to depend upon the baseline technique (LDA or SNLDA). Euclidean distance weighted based WLDA performed better than Bayes Error weighted based WLDA. WLDA achieved 10% improvement over standard LDA under *interview-interview* condition. However the marginal improvement of the performance in the *telephone-telephone* condition is likely to be due to the larger number of low session-count speaker recordings in the telephone development data, causing poorly estimated class means to reduce the quality of the estimations of the between-class scatter. This problem will be extensively investigated in future research.

Bayes Error weighted based WSNLDA outperformed the Euclidean distance weighted, because of the Bayes Error weight depends upon source variations in the within-class scatter estimation. WSNLDA system achieved 5% over SNLDA system under both mismatched conditions.

6. CONCLUSION

In this paper, we have introduced novel WLDA and WSNLDA approaches to i-vector based speaker verification system. By taking advantage of the weighted pairwise Fisher criterion, these weighted LDA techniques can take advantage of the speaker discriminative information present in the pairwise distances between classes that are not available to traditional LDA techniques. Through evaluations performed on the NIST 2008 SRE data, both WLDA and WSNLDA have shown an improvement in speaker verification performance in both matched and mismatched enrolment/verification conditions, with the best improvement in the the microphone-based *interview* conditions.

Based upon the results presented within this paper, we believe that both WLDA and WSNLDA are viable as replacement tech-

Table 1. *Speaker verification performance of weighted and non-weighted LDA, followed by WCCN, on the common set of the 2008 NIST SRE short2-short3 conditions. Column headers indicate enrolment-verification conditions.*

System	$w(d_{ij})$	interview-interview		interview-telephone		telephone-interview		telephone-telephone	
		EER	DCF	EER	DCF	EER	DCF	EER	DCF
WLDA	Euclidean	**4.14%**	**0.0199**	**5.35%**	**0.0287**	**4.89%**	**0.0213**	2.73%	**0.0128**
WLDA	Bayes Error	4.45%	0.0221	5.88%	0.0295	5.10%	0.0221	**2.72%**	0.0132
LDA	-	4.61%	0.0228	5.99%	0.0293	5.09%	0.0223	2.80%	0.0134

Table 2. *Speaker verification performance of weighted and non-weighted SNLDA, followed by WCCN, on the common set of the 2008 NIST SRE short2-short3 conditions. Column headers indicate enrolment-verification conditions.*

System	$w(d_{ij})$	interview-interview		interview-telephone		telephone-interview		telephone-telephone	
		EER	DCF	EER	DCF	EER	DCF	EER	DCF
WSNLDA	Euclidean	4.11%	0.0201	**5.34%**	0.0262	4.69%	0.0195	**2.80%**	**0.0128**
WSNLDA	Bayes Error	**4.02%**	**0.0196**	5.53%	**0.0251**	**4.41%**	**0.0184**	**2.80%**	0.0130
SNLDA	-	4.76%	0.0243	5.88%	0.0283	4.89%	0.0217	2.81%	0.0135

niques to improve the performance of LDA and SNLDA-based i-vector speaker verification.

7. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1 –1, 2010.

[2] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," *Odyssey Speaker and Language Recognition Workshop*, 2010.

[3] M. McLaren and D. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *accepted into IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011.

[4] M. Loog, R. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 7, pp. 762–766, 2001.

[5] J. Price and T. Gee, "Face recognition using direct, weighted linear discriminant analysis and modular subspaces," *Pattern Recognition*, vol. 38, no. 2, pp. 209–219, 2005.

[6] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech*, 2009.

A. WEIGHTED BETWEEN-CLASS SCATTER ESTIMATION WITH UNITY WEIGHTING FUNCTION

When weighting function $w(d_{ij})$ equals to 1, weighted between scatter estimations will converge as standard between class estimations, it can be shown as follows,

$$\mathbf{S}_b^w = \frac{1}{2N}\sum_{i=1}^{S}\sum_{j=1}^{S} n_i n_j((\bar{\mathbf{w}}_i - \bar{\mathbf{w}}) + (\bar{\mathbf{w}} - \bar{\mathbf{w}}_j)) \times((\bar{\mathbf{w}}_i - \bar{\mathbf{w}}) + (\bar{\mathbf{w}} - \bar{\mathbf{w}}_j))^T$$

$$\mathbf{S}_b^w = \frac{1}{2N}\sum_{i=1}^{S}\sum_{j=1}^{S} n_i n_j((\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^T + (\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}} - \bar{\mathbf{w}}_j)^T + (\bar{\mathbf{w}} - \bar{\mathbf{w}}_j)(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^T + (\bar{\mathbf{w}} - \bar{\mathbf{w}}_j)(\bar{\mathbf{w}} - \bar{\mathbf{w}}_j)^T)$$

Since $\sum_{i=1}^{S} \frac{n_i}{N} = 1$, we can combine the first and last outer product terms above to get

$$\mathbf{S}_b^w = \sum_{i=1}^{S} n_i(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^T + \frac{1}{2N}\sum_{i=1}^{S}\sum_{j=1}^{S} n_i n_j(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}} - \bar{\mathbf{w}}_j)^T + \frac{1}{2N}\sum_{i=1}^{S}\sum_{j=1}^{S} n_i n_j(\bar{\mathbf{w}}_j - \bar{\mathbf{w}})(\bar{\mathbf{w}} - \bar{\mathbf{w}}_i)^T$$

Examine the last two terms above, we note that $\sum_{i=1}^{S} \frac{n_i}{N}\bar{\mathbf{w}}_i = \bar{\mathbf{w}}$ and therefore $\sum_{i=1}^{S} \frac{n_i}{N}(\bar{\mathbf{w}} - \bar{\mathbf{w}}_i) = 0$. Weighted between-class scatter will converge as follows,

$$\mathbf{S}_b^w = \sum_{i=1}^{S} n_i(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^T$$