MULTI-OBJECTIVE OPTIMIZATION TRAINING OF PLDA FOR SPEAKER VERIFICATION

Liang He, Xianhong Chen, Can Xu and Jia Liu

Department of Electronic Engineering, Tsinghua University, Beijing, China.

e-mail: {heliang,chenxianhong,xucan,liuj}@mail.tsinghua.edu.cn.

ABSTRACT

Most current state-of-the-art text-independent speaker verification systems take probabilistic linear discriminant analysis (PLDA) as their backend classifiers. The parameters of PLDA are often estimated by maximizing the objective function, which focuses on increasing the value of log-likelihood function, but ignoring the distinction between speakers. In order to better distinguish speakers, we propose a multi-objective optimization training for PLDA. Experiment results show that the proposed method has more than 10% relative performance improvement in both EER and MinDCF on the NIST SRE14 i-vector challenge dataset, and about 20% relative performance improvement in EER on the MCE18 dataset.

Index Terms— probabilistic linear discriminant analysis, multi-objective optimization, i-vector, speaker verification

1. INTRODUCTION

Although there are many kinds of frontends, such as GMM/DNN i-vector [1, 2], TDNN Xvector [3], and DNN embedding [4, 5], probabilistic linear discriminant analysis (PLDA) [6, 7] is still the most popular backend for a text-independent speaker verification system.

Many researchers aim at improving the performance of PLDA [8, 9, 10, 11]. Cumani proposes a new PLDA based on i-vector's posterior distribution, where an utterance is not mapped into a single i-vector, but into a posterior distribution to improve the performance for short utterances [8]. Burget refines the PLDA scoring by adopting discriminative models, e.g. support vector machines or logistic regression [9]. Following his work, Rohdin introduces more constrains on PLDA parameters to boost system performance [10]. Inspired by the nonparametric discriminant analysis [12], Krosravani proposes a nonparametrically trained PLDA which achieves excellent performance on NIST SRE 2010 core c5 condition [11]. However, none of the above methods utilizes a discriminant way to *train* space matrix, which plays a crucial role in the PLDA modeling.

We adopt the advantages of discriminant and nonparametric methods, and propose a multi-objective optimization train-

ing for PLDA. Experiment results on the NIST SRE14 [13] and MCE18 [14] demonstrate the effectiveness of proposed methods.

The remainder of this paper is organized as follows. Section 2 reviews the simplified Gaussian probabilistic linear discriminant analysis (sGPLDA). Section 3 proposes a multi-objective optimization training for sGPLDA. Section 4 analyzes and discusses the experiment results. A conclusion is drawn in Section 5.

2. SIMPLIFIED GAUSSIAN PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS

There are many variants of PLDA, the most widely used PLDA in the field of speaker verification is the simplified Gaussian PLDA (sGPLDA) [7] for its simplicity and proven performance on the recent NIST SREs. The sGPLDA assumes that a length normalized i-vector $\mathbf{x}$ is decomposed into three parts: a global mean vector μ, a speaker space F and factor $\mathbf{h}_s$, and a ε which consists of within-class variability and residual noise.

$$\mathbf{x}_{si} = \mu + F\mathbf{h}_s + \varepsilon_{si} \qquad (1)$$

where s is the speaker index and si is the segment index of speaker s. Under the sGPLDA assumption, the speaker factor $\mathbf{h}_s$ has a standard normal prior, and $\varepsilon_{si} \sim \mathcal{N}(0, \Sigma_w)$. The model parameters $\{F, \Sigma_w\}$ and speaker factor $\mathbf{h}_s$ are iteratively optimized by maximizing the log-likelihood function f,

$$\underset{F,\Sigma_w,\mathbf{h}}{\arg\max} f = \frac{1}{\sum_{s=1}^{S} sI} \sum_{s=1}^{S} \sum_{si=1}^{sI} \\ \log \mathcal{N}(\mathbf{x}_{si}; \mu + F\mathbf{h}_s, \Sigma_w) \mathcal{N}(\mathbf{h}_s; 0, I) \qquad (2)$$

via Expectation - Maximization (EM) algorithm [15, 7]. $\mathcal{N}$ is a Gaussian distribution, S is the total speaker number and sI is the total segment number of speaker s.

3. MULTI-OBJECTIVE OPTIMIZATION TRAINING FOR SGPLDA

3.1. Motivation

Our motivation originates from the linear discriminant analysis (LDA) [16]. The LDA is to perform dimensionality reduction by analyzing within-class scatter matrix and between-class scatter matrix. The within-class scatter matrix is

$$\mathbf{S}_w = \sum_{s=1}^{S} \sum_{si=1}^{sI} (\mathbf{x}_{si} - \mu_s)(\mathbf{x}_{si} - \mu_s)^t \qquad (3)$$

and the between-class scatter matrix is

$$\mathbf{S}_b = \sum_{s=1}^{S} (\mu_s - \mu)(\mu_s - \mu)^t \qquad (4)$$

where $\mu_s = \frac{1}{sI} \sum_{si=1}^{sI} \mathbf{x}_{si}$ is the class mean, and $\mu = \frac{1}{S} \sum_{s=1}^{S} \mu_s$ is the global mean.

If we analogize sGPLDA and LDA, we will find that $\mu + F\mathbf{h}_s$ and Σ_w are equivalent to μ_s and $\mathbf{S}_w$. Finding a space F which maximizes f is similar to finding a space V which maximizes $\det(V^t \mathbf{S}_w V)^{-1}$. By these comparison, we find that the objective function f just focuses on within-class vectors, but ignores between-class vectors. Here, the within-class vectors mean that the vectors are all from the same class and the between-class vectors mean that these vectors are not from the same class. Take an extreme case for example, we only have one speaker's vectors for training. We can compute F by maximizing f and $\mathbf{S}_w$ because they only need within-class vectors, but failed to compute $\mathbf{S}_b$ because it needs between-class vectors.

Clearly, effective use of between-class statistics can further enhance the discriminant ability of designed algorithm, e.g. the space V is obtained by maximizing $\frac{\det(V^t \mathbf{S}_b V)}{\det(V^t \mathbf{S}_w V)}$ in LDA. To achieve this goal, we try to integrate between-class statistics into the sGPLDA training.

3.2. sGPLDA model for between-class vectors

For a speaker s, $\mathbf{x}_{si}$ denotes his/her i-th i-vector. Let $\bar{\mathbf{x}}_{sj}$ denote the j-th i-vector that does not belong to speaker s, $1 \leq j \leq sJ$. $\{\mathbf{x}_{si}\}$ and $\{\bar{\mathbf{x}}_{sj}\}$ constitute between-class vectors of speaker s, and we use $\mathbf{y}$ to denote them for convenience. Similar to (1) and (2), $\mathbf{y}_{sk}$ is also decomposed into three parts: a global mean vector μ, a speaker space F and factor $\mathbf{g}_s$, and a ζ which contains within-class variability of between-class vectors and residual noise.

$$\mathbf{y}_{sk} = \mu + F\mathbf{g}_s + \zeta_{sk} \qquad (5)$$

And the log-likelihood function g is

$$g = \frac{1}{\sum_{s=1}^{S} sK} \sum_{s=1}^{S} \sum_{sk=1}^{sK} \log \mathcal{N}(\mathbf{y}_{sk}; \mu + F\mathbf{g}_s, \Sigma_b)\mathcal{N}(\mathbf{g}_s; 0, I) \qquad (6)$$

Here, $sK = sI + sJ$, the factor $\mathbf{g}_s$ also has a standard normal prior, and $\zeta_{sk} \sim \mathcal{N}(0, \Sigma_b)$. We name (1) and (2) as the sGPLDA model for within-class vectors and (5) and (6) as the sGPLDA model for between-class vectors.

3.3. Multi-objective Optimization Training

The joint model parameters $\{F, \Sigma_w, \Sigma_b\}$, $\mathbf{h}_s$, and $\mathbf{g}_s$ are obtained by multi-objective optimization training, see Fig. 1. Our considerations are as follows:

1. The sGPLDA model for within-class and between-class vectors share the same speaker space F. Intuitively, the desired F is to maximize f and to minimize g at the same time. Therefore, the objective function is $\arg\max_F(\alpha f - g)$, where α is an introduced factor which balances f and g, and will be examined in the experiment section.

2. Σ_w and $\mathbf{h}_s$ are only related to f, and the objective function is $\arg\max_{\Sigma_w, \mathbf{h}_s}(f)$.

3. Σ_b and $\mathbf{g}_s$ are only related to g, and the objective function is $\arg\max_{\Sigma_b, \mathbf{g}_s}(g)$.

The parameters are obtained by EM algorithm. During the E-step, $\mathbf{h}_s$ and $\mathbf{g}_s$ are obtained by taking the derivation with f and g, respectively.

$$\mathbf{h}_s = [(sI)F^t \Sigma_w^{-1} F + I]^{-1} F^t \Sigma_w^{-1} \sum_{si=1}^{sI} (\mathbf{x}_{si} - \mu)$$

$$\mathbf{g}_s = [(sK)F^t \Sigma_b^{-1} F + I]^{-1} F^t \Sigma_b^{-1} \sum_{sk=1}^{sK} (\mathbf{y}_{sk} - \mu) \qquad (7)$$

During the M-step, F, Σ_w, and Σ_b are obtained by taking the derivation with $\alpha f - g$, f, and g, respectively.

$$F = \left(\frac{\alpha}{\sum_{s=1}^{S} sI} \sum_{s=1}^{S} \sum_{si=1}^{sI} (\mathbf{x}_{si} - \mu)\mathbf{h_s}^t \right.$$
$$\left. - \frac{1}{\sum_{s=1}^{S} sK} \sum_{s=1}^{S} \sum_{sk=1}^{sK} (\mathbf{y}_{sk} - \mu)\mathbf{g_s}^t \right)$$
$$\left(\frac{\alpha}{\sum_{s=1}^{S} sI} \sum_{s=1}^{S} (sI)(\mathbf{h}_s \mathbf{h}_s^t) \right.$$
$$\left. - \frac{1}{\sum_{s=1}^{S} sK} \sum_{s=1}^{S} (sK)(\mathbf{g}_s \mathbf{g}_s^t) \right)^{-1} \qquad (8)$$

and

$$\Sigma_w = \frac{1}{\sum_{s=1}^{S} sI} \sum_{s=1}^{S} \sum_{si=1}^{sI} (\mathbf{x}_{si} - \mu - F\mathbf{h}_s)(\mathbf{x}_{si} - \mu - F\mathbf{h}_s)^t$$

$$\Sigma_b = \frac{1}{\sum_{s=1}^{S} sK} \sum_{s=1}^{S} \sum_{sk=1}^{sK} (\mathbf{y}_{sk} - \mu - F\mathbf{g}_s)(\mathbf{y}_{sk} - \mu - F\mathbf{g}_s)^t \qquad (9)$$

The E-step and M-step are iteratively performed.

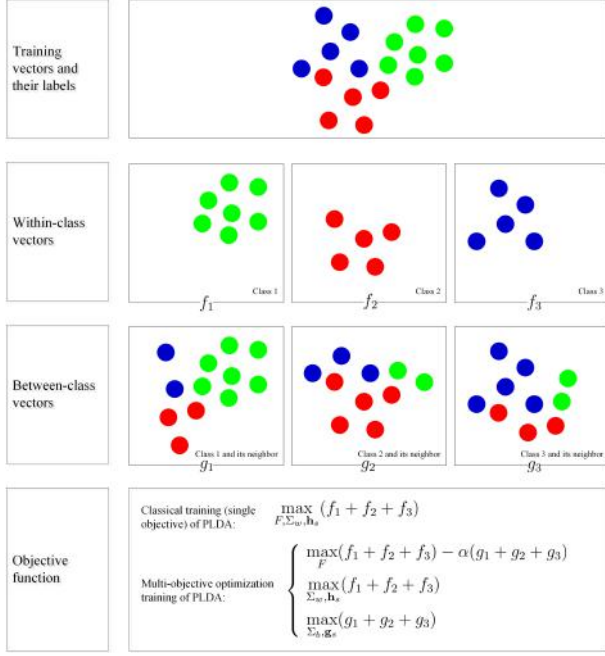**Fig. 1**. A demo of multi-objective optimization training of sGPLDA

3.4. Selection of $\bar{\mathbf{x}}_{sj}$

As stated in [17], the most challenging task in text-independent speaker verification is to discriminate easily confusable speakers. Krosravani also proposes a nonparametrically trained PLDA, in which the core idea is selecting nearest neighbor vectors during scoring [11]. Therefore, we adopt random and nearest neighbor selections to pick up $\bar{\mathbf{x}}_{sj}$. The former is used for comparison and we believe the latter is effective. The nearest neighbor selection is that for a speaker s, we calculate inner products between $\mathbf{x}_s$ and $\bar{\mathbf{x}}_{sj}$, sort them in a descending order and select the top sI $\bar{\mathbf{x}}_{sj}$.

3.5. Verification score

The scoring is also based on two-covariance model [7] and the log-likelihood ratio is

$$score(\mathbf{x}_1, \mathbf{x}_2) = \log \frac{p(\mathbf{x}_1, \mathbf{x}_2|\text{same speaker})}{p(\mathbf{x}_1, \mathbf{x}_2|\text{different speakers})}$$
$$= \mathbf{x}_1^t Q \mathbf{x}_1 + \mathbf{x}_2^t Q \mathbf{x}_2 + 2\mathbf{x}_1^t P \mathbf{x}_2 + \text{const}$$
(10)

where

$$Q = \Sigma_{tot,b}^{-1} - (\Sigma_{tot,w} - \Sigma_{ac}\Sigma_{tot,w}^{-1}\Sigma_{ac})^{-1}$$
$$P = \Sigma_{tot,w}^{-1}\Sigma_{ac}(\Sigma_{tot,w} - \Sigma_{ac}\Sigma_{tot,w}^{-1}\Sigma_{ac})^{-1}$$
$$\Sigma_{tot,w} = FF^t + \Sigma_w, \Sigma_{tot,b} = FF^t + \Sigma_b, \text{and } \Sigma_{ac} = FF^t$$
(11)

Different from [7], we use $\Sigma_{tot,b}$ instead of $\Sigma_{tot,w}$ to compute Q, because $\Sigma_{tot,b}$ is a more reasonable choice under the different speakers assumption.

4. EXPERIMENTS

4.1. NIST i-vector Machine Learning Challenge, SRE14

NIST i-vector machine learning challenge (SRE14) takes i-vectors instead of speech as input to examine the backend of speaker verification system [13]. It is gender independent, contains 1306 speaker models, 9634 test segments and 12582004 trials. Each speaker model has 5 i-vectors. The trials are randomly divided into a progress subset (40%) and an evaluation subset (60%). In addition, NIST provided a development set, containing 36572 i-vectors. All the i-vectors are 600-dimensional. We study the backend learning algorithms with development labels known. After applying LDA, traditional single objective function (SO) sGPLDA and multi-objective optimization training of sGPDLA (MO) are comparatively studied under the same condition. The dimension of LDA, SO sGPLDA, and MO sGPLDA are 250, 150, and 150, respectively. Unless otherwise specified, $\bar{\mathbf{x}}_{sj}$ is nearest selected.
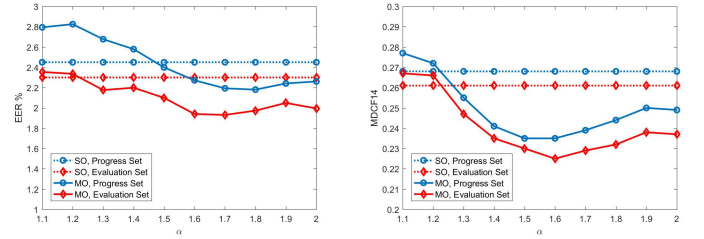

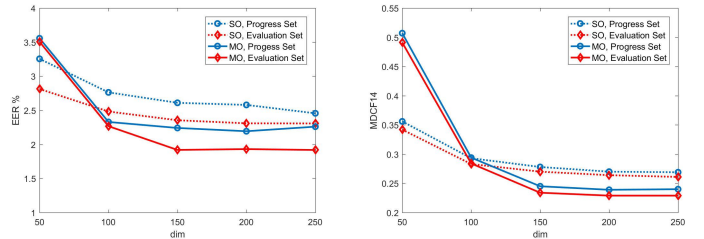**Fig. 2**. The EER and MDCF14 vary with the α

Fig. 3. The EER and MDCF14 vary with the dimension

Fig.2 shows that the EER and MDCF14 of SO/MO sG-PLDA vary with the α. It can be seen that as α changes from 1.1 to 2 with a step of 0.1, both EER and MDCF14 decrease first and then increase, which means a well balance between f and g is important for MO sGPLDA. We choose $\alpha = 1.7$ in the following experiments.

Fig.3 shows that the EER and MDCF14 of SO/MO sG-PLDA vary with the sGPLDA dimension. In most cases (150,

200, and 250), MO sGPLDA outperforms SO sGPLDA. At a low dimension (50), the performance of MO sGPLDA is decreased.
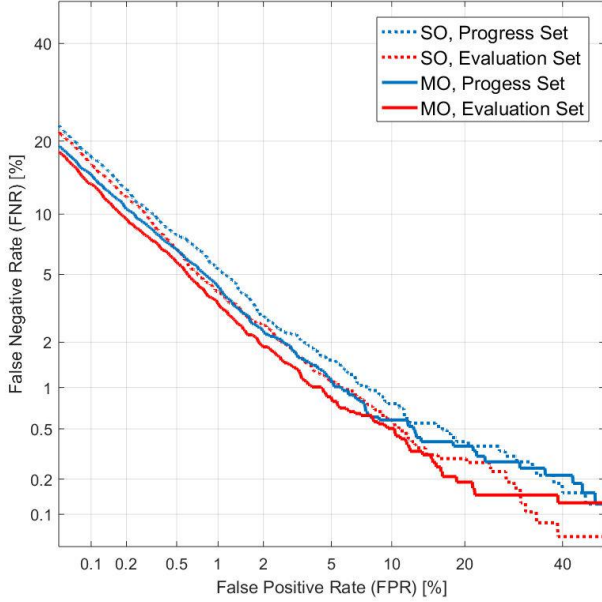

Fig. 4. The DET curves of sGPLDA on NIST SRE14

From Table 1, we can see that nearest neighbor selection is better than random selection, which is in line with our expectation. The nearest selected $\bar{x}_{sj}$ are easily confusable i-vectors. Taking them into considerations can boost system performance.

Compared with the SO sGPLDA, both EER and MDCF of proposed MO sGPLDA with nearest selection have more than 10% relative improvement on the progress set and the evaluation set of NIST SRE14, see Table 1 and Fig.4.

Table 1. Experiment results of sGPLDA on NIST SRE14.

	EER[%]	MDCF14
	Progress Set	
SO	2.45	0.268
MO,Random	3.59	0.319
MO,Nearest	**2.19**	**0.239**
	Evaluation Set	
SO	2.30	0.261
MO,Random	3.04	0.304
MO,Nearest	**1.93**	**0.229**

4.2. MCE18

The 1st Multi-target speaker detection and identification Challenge Evaluation [14] provides three i-vector sets: train-

ing, development and test sets. Each set consists of blacklist and non-blacklist (background) speakers.

For the training set, there are 3,631 blacklist speakers and 5,000 background speakers. Each blacklist speaker has 3 i-vectors, and there are 10,893 i-vectors for blacklist speakers in total. For the development set, there also 3,631 blacklist speakers and 5,000 background speakers. Each speaker has only one i-vector. The blacklist speakers of the training and development sets are the same while the background speakers are not. No information is provided about the distribution of speakers in the test set. All the i-vectors are 600 dimension. The MCE18 evaluation dataset includes the Fixed and Open conditions. In the Fixed condition, we can only use data provided by the MCE18. This limitation is removed in the Open condition. We examined the Mot sGPLDA on the Fixed condition test. Our procedure is classical, includes length normalization [15], LDA, PLDA and score normalization in turn. We use both training and development sets to train these parameters. The dimension of both LDA and PLDA is 350.

From the Table 2, we can see that, compared with the SO sGPLDA, the proposed MO sGPLDA has 19.8% and 22.0% relative improvement in the *Top S* and *Top 1* conditions on the MCE18 evaluation dataset, which further proves our assert that the parameters trained by multi-objective optimization training not only have the ability to maximize the log-likelihood function on the within vector sets, but also have the ability to distinguish the vectors which are easily mis-judged.

Table 2. Experiment results of sGPLDA on MCE18.

EER[%]	Top S	Top 1
SO	6.75	9.39
MO,Nearest	**5.41**	**7.32**

5. CONCLUSION

We propose a multi-objective optimization training for the sGPLDA. It not only focuses on increasing the log-likelihood function, but also improves the distinction ability between easily mis-judged speakers. Compared with the traditional method, the EER and MDCF of multi-objective optimized sGPLDA have 10.5% and 11.1% relative performance improvements on SRE14 progress set, and 16.2% and 12.1% relative performance improvements on SRE14 evaluation set, and the EER of multi-objective optimized sGPLDA have 19.8% and 22.0% relative performance improvements in the *Top S* and *Top 1* conditions on the MCE18 evaluation set.

This method can also be extended to other types of PLDA with proper modification. The python and matlab code for this paper can be downloaded from Github: *git clone https://github.com/sanphiee/MOT-sGPLDA-SRE14* and *git clone https://github.com/sanphiee/MOT-sGPLDA-MCE18*.

6. 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. 19, no. 4, pp. 788–798, May 2011.

[2] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech, and Signal Processing (ICASSP), 2014IEEE International Conference on*, mar. 2014, pp. 1714 – 1718.

[3] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop*, 2017, pp. 165–170.

[4] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," *Computer Science*, pp. 5115–5119, 2015.

[5] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," in *arXiv:1705.02304v1*, 2017.

[6] S. J. D Prince and J. H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[7] Botjan Vesnicer, Jerneja ganec Gros, Nikola Pavei, and Vitomir truc, "Face recognition using simplified probabilistic linear discriminant analysis," *International Journal of Advanced Robotic Systems*, vol. 9, no. 6, pp. 700–709, 2012.

[8] Sandro Cumani, Oldich Plchot, and Pietro Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7644–7648.

[9] Lukas Burget, Oldrich Plchot, Sandro Cumani, Ondej Glembek, Pavel Matejka, and Niko Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4832–4835.

[10] Johan Rohdin, Sangeeta Biswas, and Koichi Shinoda, "Constrained discriminative plda training for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1670–1674.

[11] Abbas Khosravani and Mohammad M. Homayounpour, "Nonparametrically trained plda for short duration i-vector speaker verification," *Computer Speech and Language*, vol. 52, pp. 105 – 122, 2018.

[12] Zhifeng Li, Dahua Lin, and Xiaoou Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 755–761, 2009.

[13] "The 2014 speaker recognition i-vector machine learning challenge," `https://www.nist.gov/itl/iad/mig/i-vector-machine-learning-challenge`.

[14] "Mce 2018: The 1st multi-target speaker detection and identification challenge evaluation," `http://www.mce2018.org/`.

[15] D. G. Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker verification systems," in *INTERSPEECH*, 2011, pp. 249–252.

[16] A.M. Martinez and A.C. Kak, "Pca versus lda," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 228 –233, feb 2001.

[17] L. He, X. Chen, C. Xu, J. Liu, and M. T. Johnson, "Local pairwise linear discriminant analysis for speaker verification," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1575–1579, Oct 2018.