Maximum Likelihood Linear Dimension Reduction of Heteroscedastic Feature for Robust Speaker Recognition

Suwon Shon[1], Seongkyu Mun[1], David K. Han[2], Hanseok Ko[1]
[1]School of Electrical Engineering, Korea University, Seoul, Korea
[2]Office of Naval Research, VA, USA
swshon@ispl.korea.ac.kr, skmoon@ispl.korea.ac.kr, ctmkhan@gmail.com, hsko@korea.ac.kr

Abstract

This paper analyzes heteroscedasticity in i-vector for robust forensics and surveillance speaker recognition system. Linear Discriminant Analysis (LDA), a widely-used linear dimension reduction technique, assumes that classes are homoscedastic within a same covariance. In this paper it is assumed that general speech utterances contain both homoscedastic and heteroscedastic elements. We show the validity of this assumption by employing several analyses and also demonstrate that dimension reduction using principal components is feasible. To effectively handle the presence of heteroscedastic and homoscedastic elements, we propose a fusion approach of applying both LDA and Heteroscedastic-LDA (HLDA). The experiments are conducted to show its effectiveness and compare to other methods using the telephone database of National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) 2010 extended.

1. Introduction

In surveillance or forensic application, use of vocal information has been viewed as an effective way of personal identification. For a general setting and of recognizing a person without the person's knowledge, the recognition process cannot rely on predetermined keywords or the length of speech duration. Recently, accuracy of text-independent speaker recognition has been significantly improved by the i-vector extraction paradigm [1], [2]. The i-vector approach in total variability space was first introduced in [1] and since has been considered as the state of the art in speaker verification systems. It is originated in Joint Factor Analysis (JFA) framework that consists of defining two distinct spaces: speaker and channel.

Many approaches to the subsequent step after extracting i-vector have been proposed in terms of modeling, scoring, and normalizing for reducing channel and noise effects. After the system maps the speaker utterance to an i-vector, which is a fixed length low dimensional vector, it is widely-accepted to apply i-vector length normalization [3], Linear Discriminant Analysis (LDA) [1] and Probabilistic Linear Discriminant Analysis (PLDA) [4] in sequence for improving performance. The length normalization performs centering and whitening of i-vector. Then, it scales the length of each i-vector to a unit length. It reduces any non-Gaussian behaviors of i-vector so that PLDA can use a simple Gaussian probabilistic model. For reducing computational complexity and for finding axes for better discriminant analysis, techniques such as Linear Dimension Reduction (LDR) are typically applied at this point. Among many LDR techniques, LDA is the most popular algorithm since it is effective and simple to implement. LDA finds new axes that minimize the intra-class variance caused by channel effects while also maximizing the variance between speakers. Upon dimension reduction by LDA, Prince's proposed method of modeling i-vector by PLDA has shown to be successful in speaker recognition. His PLDA infers speaker's i-vector in probabilistic sense by regarding i-vector as an observation from a probabilistic generative model. It has shown more robust performance in scoring between the target speaker and test speaker compared to those with non-probabilistic approach like Cosine Distance Similarity (CDS).

Although LDA is commonly used in many applications for LDR, it has some limitations. One such limitation is that LDA is a non-probabilistic approach. It is deterministic and thus cannot handle missing or unknown data. To deal with this problem, PLDA approach is used. The other limitation comes from the fact that the solution of LDA is optimal when classes are homoscedastic within-class covariance with the same Gaussian distribution. From this constraint of sharing the same covariance matrix, LDA fails when within-class covariance matrices are heteroscedastic. The homoscedastic assumption is too restrictive for general speech utterance because there are many discriminant information in each within-class covariance matrix. To overcome this difficulty, we propose an approach for improving speaker recognition by using a Heteroscedastic Linear Discriminant Analysis (HLDA) and its fusion.

According to Kumar and Andreou [5] and Gales [6], HLDA provides a linear transformation that can decorrelate features and reduce dimensionality while it preserves discriminant information of features. For these reasons, it is common to use HLDA in speech recognition and its combination with LDA called Smoothed HLDA (SHLDA) [7]. In the speaker recognition, Burget [8] used HLDA in feature domain to find more discriminative Mel-Frequency Cepstral Coefficients (MFCC) feature subspaces and has shown the method's effectiveness. In model domain, Glembek [9] used HLDA for orthogonalization of total variability for simplifying the speaker recognition system and reducing computational complexity.

In this paper, we explore the effect of relaxing homoscedasticity in the within covariance matrix using HLDA. Since general utterance may exhibit both homoscedastic and heteroscedastic characteristics, we propose an approach of fusing LDA and HLDA for overall performance improvement in speaker recognition. We analyze the National Institutes of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) 2010 extended database and deduce the order of dimension that has heteroscedastic discriminant information for determining the parameters in HLDA algorithm. Experiments were conducted using i-vector extracted NIST SRE 2010 extended core task of telephone-telephone condition provided by Brno University Technology (BUT) [3], [10].

The outline of the paper is as follows. First, we describe the baseline speaker recognition briefly in Section 2. Analyses in Section 3 show that the database contains heteroscedastical characteristics. Section 4 proposes an approach for linear transformation and dimensionality reduction of the database. Section 5 presents the experimental results and Section 6 concludes this paper.

2. Speaker recognition system

In classical JFA [11], a speaker utterance is represented by a supervector that consists of additive components from a speaker and a channel/session subspace. However, in total variability sense, a speaker utterance can be defined by both speaker and channel variability in a single space, i.e. total variability space [1] as in Eq. (1).

$$\mathbf{m}_s = \mathbf{m}_0 + \mathbf{T}\boldsymbol{\omega}_s \tag{1}$$

where supervector $\mathbf{m}_s$ represents the speaker utterance, $\mathbf{m}_0$ denotes the speaker and channel independent supervector, i.e. UBM supervector, $\mathbf{T}$ is a total variability matrix which is rectangular with low rank, and $\boldsymbol{\omega}_s$ is total variability factor. We refer this total variability factor as i-vector.

After extracting i-vector, length normalization and LDA are applied on it. Then, assuming the i-vector is observed from probabilistic generative model, it can be expressed as

$$\boldsymbol{\omega}_s = \boldsymbol{\mu} + \mathbf{V}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where μ is a speaker independent mean vector, $\mathbf{V}$ is basis matrix for the speaker-specific subspace, $\boldsymbol{\beta}$ is a latent identity vector having a standard normal distribution, $\boldsymbol{\varepsilon}$ is residual noise vector. The maximum likelihood point estimates of the model parameter $\{\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\Sigma}_\varepsilon\}$ are obtained from a large collection of development data using an Expectation Maximum (EM) algorithm as in [4] where $\boldsymbol{\Sigma}_\varepsilon$ represents full covariance residual noise.

3. Heteroscedasticity in i-vector

For analyzing heteroscedasticity in speaker recognition, we use an i-vector extracted from NIST SRE 04, 05, 06 and Switchboard II phases 2 and 3, Switchboard cellular parts 1 and 2 database provided by BUT [3, 10]. For 400 dimensional i-vector extractions, a full-covariance gender-independent UBM with 2048 mixture was used. Then, NIST SRE 04 and 05 telephone data were used for training UBM using a 20ms short-time Gaussianized MFCC plus delta and double-delta features. A gender-dependent i-vector extractor was trained from the telephone data of the NIST SRE 04, 05, 06, Switchboard and Fisher.

The covariance matrix for j-th speaker i-vector within class is

$$\Sigma^{(j)} = \frac{1}{N_j}\sum_{i=1}^{N_j}(\boldsymbol{\omega}_i^{(j)} - \bar{\boldsymbol{\omega}}^{(j)})(\boldsymbol{\omega}_i^{(j)} - \bar{\boldsymbol{\omega}}^{(j)})^T \cdot \tag{3}$$

N_j is the total utterance number of j-th speaker. LDA regard the speaker i-vector within class covariance matrix as same within class covariance matrix $\mathbf{S}_w$.

$$\mathbf{S}_w = \frac{1}{N}\sum_{j=1}^{S} N_j \Sigma^{(j)} \tag{4}$$

N is the number of total utterance.

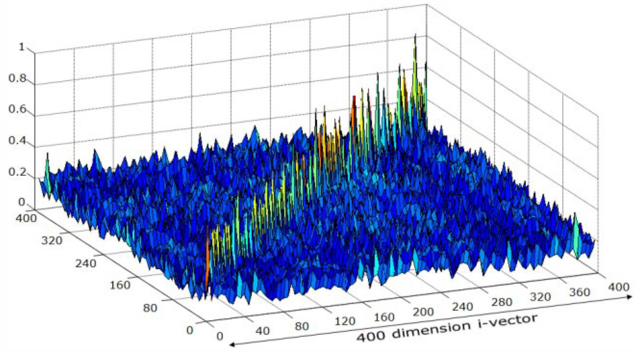


Figure 1. Variance matrix of each speaker class i-vector covariance in male telephone database of NIST SRE 04, 05, 06 and switchboard

For validating the heteroscedasticity, we obtain the variance matrix of speaker i-vector within class covariance matrix as

$$\boldsymbol{\sigma}_{\mathbf{S}_w} = \sum_{j=1}^{N_j} \frac{1}{N_j} (\Sigma^{(j)} - \mathbf{S}_w)(\Sigma^{(j)} - \mathbf{S}_w)^T \tag{5}$$

Figure 1 is developed using Eq. (5). As expected, it is clear from Figure 1 that each speaker i-vector within class covariance is not homoscedastic. Especially, the diagonal components have higher variance than the off-diagonal components. The variance has risen from many factors like noise or difference of vocal characteristics among speakers. Major contributing factor we discovered, however, was from the *difference of database length*. Hasan et. al. [12] found that the deviation of the i-vector depends on the utterance duration.

They observed that as duration becomes shorter, the average deviation increases. Since there are many different durations in the development data (NIST SRE 04, 05, 06 and switchboard), it cannot have equal covariance matrix. In statistical sense, this is reasonable since the longer duration equates to larger sample size, thus the smaller variance.

From these analyses, speaker i-vector within class covariance contains relevant information regarding speaker recognition. Therefore, by using this variance matrix $\boldsymbol{\sigma}_{S_w}$, the order of dimension needed for covering the heteroscedastic discriminant information can be determined. We perform eigenvalue decomposition to variance matrix $\boldsymbol{\sigma}_{S_w}$ for looking into eigenvalue energy distribution. Figure 2 shows sorted eigenvalues of the variance matrix $\boldsymbol{\sigma}_{S_w}$.

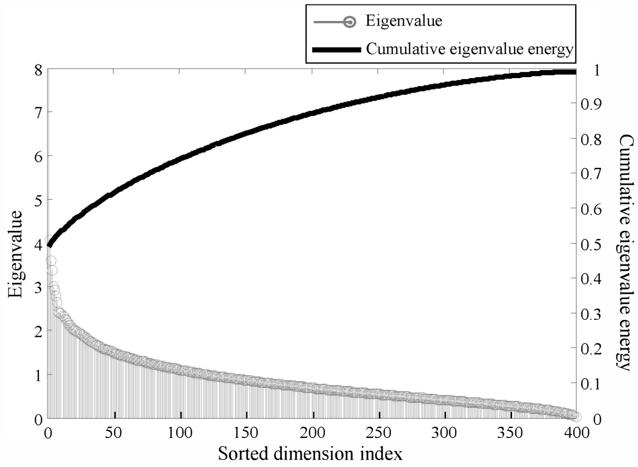


Figure 2. Sorted eigenvalue decomposed from variance matrix of speaker i-vector within class covariance and its cumulative energy

As shown, it is observed that almost 90% of the energy is concentrated on the first 200 dimensions. It tells us that only about half the dimension is needed for containing the heteroscedastic discriminant information. This analysis result will be shown validated in the experiments in Section 5.

4. HLDA Transform of i-vector and unified HLDA transform

From the analysis in Section 3, it has been shown that homoscedastic assumption may not be effective in dealing with the utterances of heteroscedasticity. Therefore, HLDA is applied here for heteroscedastic speaker i-vector within class covariance [5]. Using the HLDA transformation matrix, we can obtain i-vector $\hat{\mathbf{w}}$ as

$$\hat{\mathbf{w}} = \mathbf{A}\mathbf{w} = \begin{bmatrix} \mathbf{A}_{[p]}\mathbf{w} \\ \mathbf{A}_{[N-p]}\mathbf{w} \end{bmatrix} \tag{6}$$

where **A** is *M*-by-*N* matrix consisting of first *p* rows for the useful dimensions $\mathbf{A}_{[p]}$ and remaining *N*-*p* rows for the nuisance dimensions $\mathbf{A}_{[N\text{-}p]}$.

To find the optimal HLDA transformation matrix **A**, Kumar [5] used a standard nonlinear optimization technique. However, his technique required a high computational complexity and large memory capacity. Gales proposed a simple iterative optimization scheme that is based on the EM algorithm [6]. Hence, for the HLDA implementation, Gales' optimization approach was adopted in our method.

Although the speaker i-vector within class covariance matrices are mostly heteroscedastic, from the analysis, it has also been shown that LDA with homoscedastic assumption may work in part [1], [10]. It can be inferred that general speech utterances may have both homoscedastic and heteroscedastic elements. Moreover, LDA reduces noise in the covariance matrix of each within class covariance as in Eq. (4). The reason is that higher the number of classes, fewer the available feature data for each class. Thus, its covariance becomes noisier [7]. From these reasons, we propose a fusion of LDA and HLDA by a transform matrix called Unified HLDA (UHLDA). The proposed fusion ensures the UHLDA to remain in the most principal subspace of LDA and HLDA as

$$\mathbf{C} = [\mathbf{A}_{q/2} \quad \mathbf{W}_{q/2}] \tag{7}$$

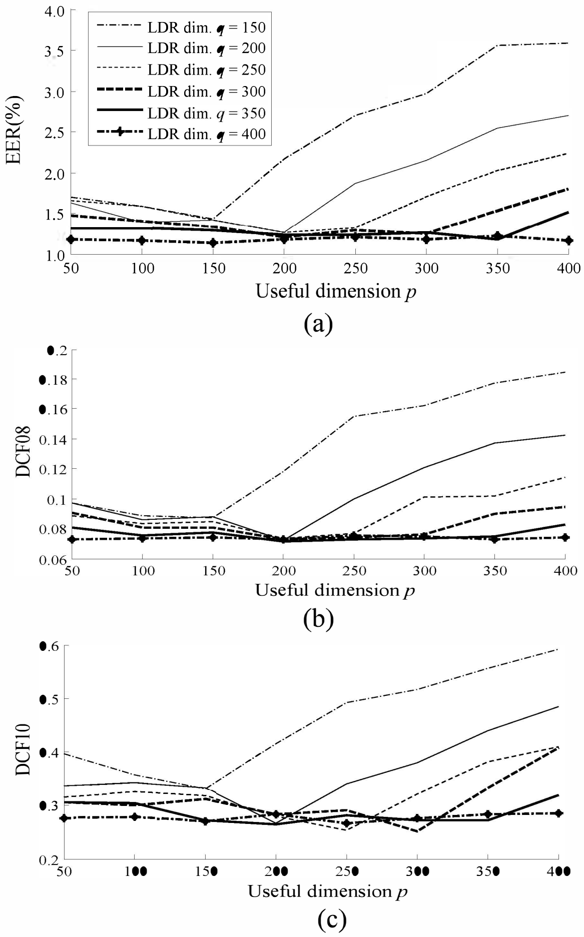
where

$$\mathbf{A}_{q/2} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,q/2} \\ \vdots & \ddots & \vdots \\ a_{M,1} & \cdots & a_{M,q/2} \end{bmatrix} \qquad \mathbf{W}_{q/2} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,q/2} \\ \vdots & \ddots & \vdots \\ w_{M,1} & \cdots & w_{M,q/2} \end{bmatrix}.$$

Note that q is the LDR dimension, $\mathbf{C}$ is M-by-q matrix and $\mathbf{W}$ is LDA transform matrix with M-by-N. $\mathbf{A}_{q/2}$ and $\mathbf{W}_{q/2}$ are each subspace of $\mathbf{A}$ and $\mathbf{W}$ respectively with $q/2$ rows. $a_{m,n}$ and $w_{m,n}$ are the components of $\mathbf{A}$ and $\mathbf{W}$ matrices at m row and n column.

5. Experiments

The experiments were conducted on the NIST SRE 2010 extended core task and telephone-telephone condition (i.e. common evaluation condition 5). As mentioned in Section 3 we use i-vector which is provided by BUT. All i-vectors are normalized by whitening and scaling the length of each i-vector to a unit length [3]. We reduced the i-vector from 400 dimensions to q dimension using LDR technique such as LDA or HLDA. After dimensional reduction, the speaker and session dependent i-vector distribution is modeled using PLDA with 150 dimension eigenvoices and a full covariance noise matrix. We measured the system performance in terms of Equal Error Rate (EER), Decision Cost Function (DCF) defined in 2008 (DCF08) and 2010 (DCF10) in NIST SRE plan.



(a)

(b)

(c)

Figure 3. Performance evaluation using HLDA for LDR technique followed by PLDA corresponding useful dimension p and LDR dimension q with respect to (a) EER (b) DCF08 (c) DCF10

Table 1. Performance evaluation when useful dimension p and LDR dimension q is 200

System name	Female			Male		
	EER (%)	dcf08	dcf10	EER (%)	dcf08	dcf10
PLDA	5.28	0.2636	0.6649	3.60	0.1897	0.5727
LDA-PLDA	2.58	0.1254	0.4048	1.33	0.0782	0.2802
HLDA-PLDA	2.51	0.1220	0.4009	1.26	0.0721	0.2710
SHLDA-PLDA	2.35	0.1160	**0.3803**	1.27	0.0740	0.2753
UHLDA-PLDA	**2.32**	**0.1125**	0.3818	**1.23**	**0.0693**	**0.2573**

Table 2. Performance evaluation when useful dimension p and LDR dimension q is 400

System name	Female			Male		
	EER (%)	dcf08	dcf10	EER (%)	dcf08	dcf10
PLDA	2.43	0.1165	0.4019	1.19	0.0773	0.2854
LDA-PLDA	2.39	0.1172	0.4028	1.23	0.0764	0.2889
HLDA-PLDA	2.37	0.1173	0.3846	**1.13**	0.0740	0.2760
SHLDA-PLDA	2.34	0.1189	0.3854	1.14	0.0733	0.2870
UHLDA-PLDA	**2.32**	**0.1156**	**0.3723**	1.14	**0.0718**	**0.2612**

The first experiment is a performance evaluation of male gender trials using HLDA corresponding useful dimension p and LDR dimension q. The goal here is to determine the useful dimension p for the best performance.

Figure 3 shows the performance of a speaker recognition system using HLDA for LDR technique. The three indices of performance were measured while useful dimension p and LDR dimension q were varied from 50 to 400 and 150 to 400 respectively in 50 step increments as shown in Figure 3 (a)-(c). The following observations can be drawn from these tests:

1) Useful dimension p must be same or under LDR dimension q.

2) Higher LDR dimension is better.

3) The performance saturates when useful dimension p is higher than half dimension (=200). It validates that the analysis result at section 3.

Using these information, we can determine parameter p and q for the best performance in a certain recognition environment.

In the second experiment, we compared the performance of various speaker recognition systems for both genders when the parameters, useful dimension p and LDR dimension q, were set at 200 as shown Table 1. Other settings are exactly the same with the first experiment. In addition, we obtained the performance when the parameters *p* and *q* are 400 for using all of the dimensions regardless of computational load as in Table 2. The eigenvoice dimension for PLDA is the same as the first experiment. The smoothing factor for SHLDA is 0.75 for the best result [7]. For the system name in Tables 1 and 2, PLDA means that the system does not use LDR technique and only uses first q dimension i-vector. From the two results, it is apparent that the system using HLDA demonstrated the best result. Compared to the LDA-PLDA system when the dimension is 200, HLDA-PLDA system showed improved performance in all three metrics. The systems using SHLDA and the proposed UHLDA showed improvements in performance compared to that of the HLDA. Especially, SHLDA-PLDA shows better performance at EER, whereas UHLDA-PLDA shows better at DCF08 and DCF10.

6. Conclusion

In this paper, we proposed to apply HLDA as an LDR technique and a fusion approach UHLDA. In-depth analyses of heteroscedasticity in i-vector were conducted for robust speaker recognition. Through the analyses in i-vector, we found performance improvement and validated it from the experiments. In addition, the fusion approach in UHLDA was verified.

The reason for the heteroscedasticity in i-vector is the difference of database durations. Therefore, our future work will be to explore how heteroscedastic approach influences in speaker recognition system when the duration mismatch condition occurs.

Acknowledgements

This research was supported in part by the BK21Plus Program and in part by Korea University Grant

References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[2] P. Kenny, “Bayesian speaker verification with heavy tailed priors,” *Odyssey Speker Lang. Recognit. Work. Brno, Czech Repub.*, 2010.

[3] D. Garcia-romero and C. Y. Espy-wilson, “Analysis of i-vector Length Normalization in Speaker Recognition Systems.,” in *Interspeech*, 2011, no. August, pp. 249–252.

[4] S. Prince and J. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *International Conference on Computer Vision*, 2007, pp. 1–8.

[5] N. Kumar and A. Andreou, “Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition,” Diss. Johns Hopkins University, 1997.

[6] M. Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 272–281, 1999.

[7] L. Burget, “Combination of speech features using smoothed heteroscedastic linear discriminant analysis.,” in *Interspeech*, 2004, pp. 2549–2552.

[8] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, “Analysis of feature extraction and channel compensation in a GMM speaker recognition system,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 1979–1986, 2007.

[9] O. Glembek and L. Burget, “Simplification and optimization of i-vector extraction,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 4516–4519.

[10] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, “Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 4828–4831.

[11] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms,” *CRIM, Montr. CRIM-06/08-13*, pp. 1–17, 2005.

[12] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, “Duration mismatch compensation for i-vector based speaker recognition systems,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7663–7667.