Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

A Double Joint Bayesian Approach for J-Vector Based Text-dependent Speaker Verification

Ziqiang Shi, Mengjiao Wang, Liu Liu, Huibin Lin, Rujie Liu

Fujitsu Research and Development Center, Beijing, China
shiziqiang@fujitsu.com.cn

Abstract

J-vector has been proved to be very effective in text-dependent speaker verification with short-duration speech. However, the current state-of-the-art back-end classifiers, e.g. joint Bayesian model, cannot make full use of such deep features. In this paper, we generalize the standard joint Bayesian approach to model the multi-faceted information in the j-vector explicitly and jointly. In our generalization, the j-vector was modeled as a result derived by a generative Double Joint Bayesian (DoJoBa) model, which contains several kinds of latent variables. With DoJoBa, we are able to explicitly build a model that can combine multiple heterogeneous information from the j-vectors. In verification step, we calculated the likelihood to describe whether the two j-vectors having consistent labels or not. On the public RSR2015 data corpus, the experimental results showed that our approach can achieve 0.02% EER and 0.02% EER for impostor wrong and impostor correct cases respectively.

1. Introduction

As opposed to text-independent speaker verification, where the speech content is unconstrained, text-dependent speaker verification systems are more favorable for security applications since they showed higher accuracy on short-duration sessions [1, 2].

The previous methods regarding text-dependent speaker verification can be grouped into two categories. The first category is based on the traditional state-of-the-art GMM-UBM and i-vector approach, which may not work well in this case [3, 1, 4]. In the second category, deep models are transferred to speaker verification: deep neural network (DNN) is used to estimate the frame posterior probabilities [5]; DNN as a feature extractor for the utterance representation [6]; Zeinali et al. [7] have shown that using bottle-neck DNN features (BN) concatenated to other acoustic features outperformed the DNN method for text-dependent speaker verification; multi-task learning jointly learns both speaker identity and text information [8].

This paper is based on two works: one is of Chen et al. [8], in which the j-vector was introduced as a kind of more compact representation for text dependent utterances, and the classic probability linear discriminant analysis (PLDA) was used as the back-end classifier [9, 10, 11]; the other is of Chen et al. [12], in which the state-of-the-art joint Bayesian analysis is proposed to model two facial images jointly with an appropriate prior that considers intra- and extra-personal variations over the image pairs. However, the standard joint Bayesian model only considers one single label, but in practice the extracted features are always associated with several labels, for example when using multi-task learned networks as feature extractor to obtain j-vectors [8]. Since j-vectors potentially have different kinds of labels, the latent text variable is no longer dependent on the current label only, but it rather depends on a separate text label. This means for j-vector there are two latent variables, namely speaker and text variables, that cannot be modeled independently using j-vectors, and both variables are tied across all samples that sharing a certain label.

In order to improve j-vector modeling, we propose a generalization of the standard joint Bayesian analysis [12, 13] called Double Joint[1] Bayesian (DoJoBa), which can explicitly and jointly model the multi-view information from samples, such as certain individual saying some text content. The relationship between DoJoBa and standard joint Bayesian is analogous to that between joint factor analysis and factor analysis. DoJoBa is also related to the work of Shi et al. [14], in which a joint PLDA is proposed for j-vector verification. One of the most important advantages of DoJoBa compared to joint PLDA, is that DoJoBa can learn the appropriate dimensionality (or the number of columns) of the low-rank speaker subspace and phrase subspaces without user tuning.

The remainder of this paper is organized as follows: Section 2 reviews the standard j-vector/joint Bayesian system. Section 3 describes the DoJoBa approach. The detailed experimental results and comparisons are presented in Section 4 and the whole work is summarized in Section 5.

2. Baseline j-vector/joint Bayesian model

The standard j-vector representation [8] and the joint Bayesian model [12] are used as baselines in this work. This section gives a brief review of these baselines.

2.1. J-vector extraction

Chen et al. [8] proposed a method to train a DNN to make classifications for both speaker and phrase by minimizing a total loss function consisting of a sum of two cross-entropy losses as shown in Fig. 1 - one related to the speaker label and the other to the text label. Once training is completed, the output layer is removed, and the rest of the neural network is used to extract speaker-phrase joint features. Each frame of an utterance is forward propagated through the network, and the output activations of all the frames are averaged to form an utterance-level feature vector called j-vector. The enrollment speaker models are formed by averaging the j-vectors corresponding to the enrollment recordings.

[1] For the "double joint" term, the first "joint" is for modeling the multi-view information jointly, e.g. text and identity in j-vector; while the second "joint" is for joint distribution of two features, e.g. target and test j-vectors.

10.21437/Odyssey.2018-51

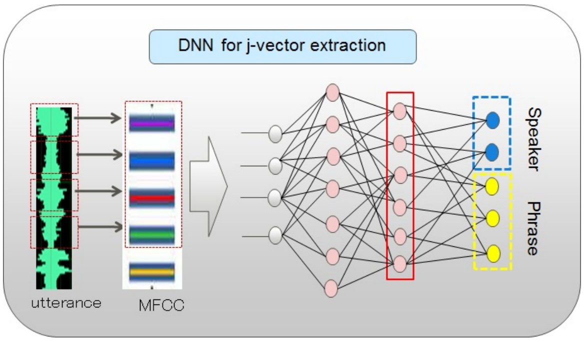

Figure 1: Multi-task joint learning DNN as j-vector extractor.

2.2. The joint Bayesian model

For the back-end, the state-of-the-art joint Bayesian model [12] is employed as a classifier for speaker verification. For simplicity of notation, joint Bayesian model with only single speaker label is used here as an example. Assume there are I speakers each with H_i sessions. We denote the the feature vector of the j'th session of the i'th speaker by x_{ij}. Then the joint Bayesian approach models data generation using the following equation:

$$x_{ij} = \mu + z_i + \epsilon_{ij}$$

where x_{ij} is a feature vector, and z_i and ϵ_{ij} are defined to be Gaussian with diagonal covariance Σ_z and Σ_ϵ respectively.

The parameters $\theta = \{\mu, \Sigma_z, \Sigma_\epsilon\}$ of this joint Bayesian model can be estimated using the Expectation Maximization (EM) [15, 12] algorithm. With the learned joint Bayesian model, given a test x_t and an enrolled model x_s, the likelihood ratio score is

$$l(x_t, x_s) = \frac{P(x_t, x_s|\text{same-speaker})}{P(x_t, x_s|\text{different-speakers})}.$$

3. Double joint Bayesian model

3.1. Motivation

The standard joint Bayesian approach cannot properly deal with j-vectors that jointly belong to certain speaker and certain phrase at the same time. For j-vector, it is noted that we need to define the joint Bayesian latent variable z_i as the joint variable considering both speaker and phrase information. This means the latent variable z_i is dependent on both a speaker identity and a phrase label. In this work we try to separate the z_i into two independent latent variables - one related to the speaker identity information and the other to the phrase. This intuitive idea results in the following DoJoBa.

In this section, we propose an effective model to describe the j-vector as resulting from a generative model which incorporates both intra-speaker/phrase and inter-speaker/phrase variation.

3.2. Generative model

We assume that the training data is obtained from I speakers saying J phrases each with H_{ij} sessions. We denote the j-vector of the k'th session of the i'th speaker saying j'th phrase by x_{ijk}. We model the text dependent feature generation by the process:

$$x_{ijk} = \mu + u_i + v_j + \epsilon_{ijk}. \tag{1}$$

The model comprises two parts: 1, the signal component $\mu+u_i+v_j$ which depends only on the speaker and phrase, rather than on the particular feature vector (i.e. there is no dependence on k); 2, the noise component ϵ_{ijk} which is different for every feature vector of the speaker/phrase and represents within-speaker/phrase noise. The term μ represents the overall mean of the training vectors. Remaining unexplained data variation is explained by the residual noise term ϵ_{ijk} which is defined to be Gaussian with diagonal covariance Σ_ϵ. The latent variables u_i and v_j are defined to be Gaussian with diagonal covariance Σ_u and Σ_v respectively, and are particularly important in real application, as these represents the identity of the speaker i and the content of the text j respectively.

Formally the model can be described in terms of conditional probabilities

$$\begin{aligned} p(x_{ijk}|u_i, v_j, \theta) &= \mathcal{N}(x_{ijk}|\mu + u_i + v_j, \Sigma_\epsilon), \\ p(u_i) &= \mathcal{N}(u_i|0, \Sigma_u), \\ p(v_j) &= \mathcal{N}(v_j|0, \Sigma_v). \end{aligned}$$

where $\mathcal{N}(x|\mu, \Sigma)$ represents a Gaussian in x with mean μ and covariance Σ. Here it's worth to notice that the mathematical relationship between DoJoBa and joint Bayesian [12] is analogous (not exactly) to that between joint PLDA [14] and PLDA [16]. Compared to joint PLDA, DoJoBa allows the data to determine the appropriate dimensionality of the low-rank speaker and text subspaces for maximal discrimination, as opposed to requiring heuristic manual selections.

Let $X = \{x_{ijk} \in \mathbb{R}^D : i = 1, ..., I; j = 1, ..., J; k = 1, ..., H_{ij}\}$, $x_{ij} = \{x_{ijk} : k = 1, ..., H_{ij}\}$, and $x_i = \{x_{ijk} : j = 1, ..., J; k = 1, ..., H_{ij}\}$. In order to maximize the likelihood of data set X with respect to parameters $\theta = \{\mu, \Sigma_u, \Sigma_v, \Sigma_\epsilon\}$, the classical EM algorithm [15] is employed.

3.3. EM formulation

The auxiliary function for EM is

$$\begin{aligned} &Q(\theta|\theta_t) = \mathrm{E}_{U,V|X,\theta_t}[\log p(X, U, V|\theta)] \\ &= \mathrm{E}_{U,V|X,\theta_t}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} \log[p(x_{ijk}|u_i, v_j, \theta)p(u_i, v_j)]\right\} \end{aligned}$$

By maximizing the auxiliary function, we obtain the following EM formulations.

E steps:we need to calculate the expectations $\mathrm{E}_{U|X,\theta_t}[u_i]$, $\mathrm{E}_{V|X,\theta_t}[v_j]$, $\mathrm{E}_{U|X,\theta_t}[u_i u_i^T]$, $\mathrm{E}_{V|X,\theta_t}[v_j v_j^T]$, and $\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T]$. Indeed we have

$$\mathrm{E}_{U|X,\theta_t}[u_i] = \left(\Sigma_u^{-1} + \Sigma_\epsilon^{-1}\sum_{j=1}^{J} H_{ij}\right)^{-1} \Sigma_\epsilon^{-1}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}(x_{ijk} - \mu - v_j). \tag{2}$$

and

$$\mathrm{E}_{U|X,\theta_t}[u_i u_i^T] = \left(\Sigma_u^{-1} + \Sigma_\epsilon^{-1}\sum_{j=1}^{J} H_{ij}\right)^{-1} + \mathrm{E}_{U|X,\theta_t}[u_i]\mathrm{E}_{U|X,\theta_t}[u_i]^T. \tag{3}$$

It is almost the similar equations for $\mathrm{E}_{V|X,\theta_t}[v_j]$ and

$\mathrm{E}_{V|X,\theta_t}[v_j v_j^T]$. For $\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T]$, we have

$$\mathrm{E}_{U,V|X,\theta_t}\left\{\begin{bmatrix} u_i u_i^T & u_i v_j^T \\ v_j u_i^T & v_j v_j^T \end{bmatrix}\right\} = \tag{4}$$

$$\left(\mathbf{diag}[\Sigma_u^{-1}, \Sigma_v^{-1}] + H_{ij}\mathbf{B}^T\Sigma_\epsilon^{-1}\mathbf{B}\right)^{-1}$$

$$+\mathrm{E}_{U,V|X,\theta_t}\left\{\begin{bmatrix} u_i \\ v_j \end{bmatrix}\right\}\mathrm{E}_{U,V|X,\theta_t}\left\{\begin{bmatrix} u_i \\ v_j \end{bmatrix}\right\}^T$$

where $\mathbf{B} = \begin{bmatrix}\mathbf{I} & \mathbf{I}\end{bmatrix}$ and

$$\mathrm{E}_{U,V|X,\theta_t}\left\{\begin{bmatrix} u_i \\ v_j \end{bmatrix}\right\} =$$

$$\left(\mathbf{diag}[\Sigma_u^{-1}, \Sigma_v^{-1}] + H_{ij}\mathbf{B}^T\Sigma_\epsilon^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\Sigma_\epsilon^{-1}\sum_{k=1}^{H_{ij}}(x_{ijk} - \mu).$$

M steps: we update the values of the parameters $\theta = \{\mu, \Sigma_u, \Sigma_v, \Sigma_\epsilon\}$ and have

$$\Sigma_u = \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} 1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}\mathrm{E}_{U|X,\theta_t}[u_i u_i^T],$$

$$\Sigma_v = \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} 1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}\mathrm{E}_{V|X,\theta_t}[v_j v_j^T],$$

$$\begin{aligned}\Sigma_\epsilon \;&=\; \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} 1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}\{(x_{ij} - \mu)(x_{ijk} - \mu)^T \\ &-\; 2(x_{ijk} - \mu)[\mathrm{E}_{U|X,\theta_t}[u_i]^T + \mathrm{E}_{V|X,\theta_t}[v_i]^T] \\ &+\; \left(\mathrm{E}_{U|X,\theta_t}[u_i u_i^T] + 2\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T] + \mathrm{E}_{V|X,\theta_t}[v_j v_j^T]\right)\},\end{aligned}$$

and

$$\mu = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} x_{ijk}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} 1}.$$

The expectation terms $\mathrm{E}_{U|X,\theta_t}[u_i]$, $\mathrm{E}_{V|X,\theta_t}[v_j]$, $\mathrm{E}_{U|X,\theta_t}[u_i u_i^T]$, $\mathrm{E}_{V|X,\theta_t}[v_j v_j^T]$, and $\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T]$ can be extracted from Equations (2), (3) and (4).

3.4. Likelihood Ratio Scores

We treat the verification as a kind of hypothesis testing problem with the null hypothesis $\mathcal{H}_0$ where two j-vectors have the same speaker and phrase variables u_i and v_j and the alternative hypothesis $\mathcal{H}_1$ where they do not (there are three cases: different underlying u_i variable with same v_j variable in model $\mathcal{M}_1$, same u_i variable with different v_j variables in model $\mathcal{M}_2$, or different underlying u_i variables with different v_j variables in model $\mathcal{M}_3$, as shown in Fig. 2). Given a test j-vector x_t and an enrolled j-vector x_s, and let a priori probability of the models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ be $p_1 = P(\mathcal{M}_1|\mathcal{H}_1)$, $p_2 = P(\mathcal{M}_2|\mathcal{H}_1)$, $p_3 = P(\mathcal{M}_3|\mathcal{H}_1)$, respectively, then the likelihood ratio score is

$$l(x_t, x_s) = \frac{P(x_t, x_s|\mathcal{H}_0)}{P(x_t, x_s|\mathcal{H}_1)}$$

$$= \frac{\int\int p(x_t, x_s|u_1, v_1, \theta)p(u_1)p(v_1)du_1 dv_1}{\mathbf{X}}$$

$$= \frac{\mathcal{N}(\begin{bmatrix} x_t \\ x_s \end{bmatrix} | \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_u + \Sigma_v + \Sigma_\epsilon & \Sigma_u + \Sigma_v \\ \Sigma_u + \Sigma_v & \Sigma_u + \Sigma_v + \Sigma_\epsilon \end{bmatrix})}{\mathbf{X}},$$

where

$$\begin{aligned}&\mathbf{X} = P(x_t, x_s|\mathcal{H}_1) = P(x_t, x_s|\mathcal{M}_1)P(\mathcal{M}_1|\mathcal{H}_1) \\ &+P(x_t, x_s|\mathcal{M}_2)P(\mathcal{M}_2|\mathcal{H}_1) + P(x_t, x_s|\mathcal{M}_3)P(\mathcal{M}_3|\mathcal{H}_1) \\ &=p_1\int\int\int p(x_t, x_s, u_1, u_2, v_1|\theta)du_1 du_2 dv_1 \\ &+p_2\int\int\int p(x_t, x_s, u_1, v_1, v_2|\theta)du_1 dv_1 dv_2 \\ &+p_3\int\int p(x_t, u_1, v_1|\theta)du_1 dv_1 \int\int p(x_s, u_2, v_2|\theta)du_2 dv_2 \\ &=p_1\mathcal{N}(\begin{bmatrix} x_t \\ x_s \end{bmatrix} | \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_u + \Sigma_v + \Sigma_\epsilon & \Sigma_v \\ \Sigma_v & \Sigma_u + \Sigma_v + \Sigma_\epsilon \end{bmatrix}) \\ &+p_2\mathcal{N}(\begin{bmatrix} x_t \\ x_s \end{bmatrix} | \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_u + \Sigma_v + \Sigma_\epsilon & \Sigma_u \\ \Sigma_u & \Sigma_u + \Sigma_v + \Sigma_\epsilon \end{bmatrix}) \\ &+p_3\mathcal{N}(x_t|\mu, \Sigma_u + \Sigma_v + \Sigma_\epsilon)\mathcal{N}(x_s|\mu, \Sigma_u + \Sigma_v + \Sigma_\epsilon).\end{aligned}$$

Notice that we do not calculate a points estimate of the hidden variable as in the standard joint Bayesian model [12]. Instead we compute the probability that the two multi-label vectors had the same hidden variables, regardless of what this actual latent variable was.

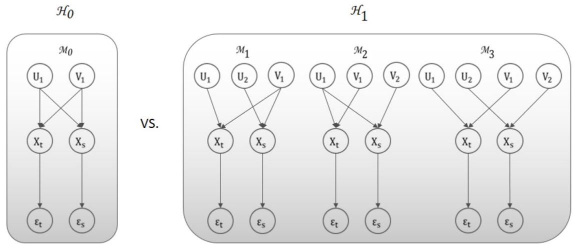

Figure 2: J-vectors x_t and x_s match under the null hypothesis $\mathcal{H}_1$, and they do not match under hypothesis $\mathcal{H}_0$.

4. Experiments

In this section, we describe the experimental setup and results for the proposed method on the public RSR2015 English corpus [1] and our internal Huiting202 Chinese Mandarin database collected by the Huiting Techonogly[2].

4.1. Experimental setup

RSR2015 corpus [1] was released by the Technology Institute for Infocomm Research and A*STAR, is used to evaluate the performance of different speaker verification systems. In this work, we follow the setup of [17], the part I of RSR2015 is used for the testing of DoJoBa. The background and development data of RSR2015 part I are merged as new background data to train the j-vector extractor.

Our internal gender balanced Huiting202 database is designed for local applications. It contains 202 speakers reading 20 different phrases, 20 sessions each phrase. All speech files were recorded with a sampling frequency of 16kHz. 132 randomly selected speakers are used for training the background multi-task learned DNN, and the remaining 70 speakers were used for enrollment and evaluation.

In this work, 39-dimensional Mel-frequency cepstral coefficients (13 MFCC plus log energy, delta and acceleration

[2] http://huitingtech.com/

coefficients) are extracted and normalized using utterance-level mean and variance normalization. The input is stacked normalized MFCCs from 11 frames (5 frames from each side of the current frame) and the current frame. The DNN has 6 hidden layers (with sigmoid activation function) of 2048 nodes each. During the background model development stage, the DNN was trained by the strategy of pre-training with Restricted Boltzmann Machine (RBM) [18] and fine tuning with SGD using cross-entropy criterion. Once the DNN is trained, the j-vector can be extracted during the enrollment and evaluation stages.

4.2. Results and discussion

Four systems are evaluated and compared across above conditions:

- **j-vector**: the standard j-vector system with cosine similarity [8].
- **joint Bayesian**: the j-vector system with classic joint Bayesian in [12].
- **jPLDA**: joint PLDA system described in [14] with j-vector.
- **DoJoBa**: double joint Bayesian system described in Section 3 with j-vector.

When evaluation a speaker three utterances of the same text were used for enrollment. The task concerns on both the phrase content and speaker identity. Since the task considers both text content and speaker identity, there are three types of nontarget trials: an impostor pronouncing wrong lexical content (impostor wrong, IW); a target speaker pronouncing wrong lexical content (target wrong, TW); the imposter pronouncing correct lexical content (impostor correct, IC).

The joint Bayesian, jPLDA, and DoJoBa models are trained using the j-vectors. The class defined in all models is the multi-task label of both the speaker and phrase. For each test session the j-vector is extracted using the same process and then the log likelihood from joint Bayesian, jPLDA, and DoJoBa are used to distinguish among different models. The number of principal components is set to 100 and then the joint Bayesian model is estimated with 10 iterations; the speaker and the phrase subspace dimensions of jPLDA and DoJoBa are both set to 100 regarding of fair comparisons and the jPLDA and DoJoBa model are also trained with 10 iterations.

Tables 1 and 2 compare the performances of all above-mentioned systems in terms of equal error rate (EER) for the three types of nontarget trials. As shown by the experimental results shown in the tables, DoJoBa is superior to the standard joint Bayesian and jPLDA, regardless of the test database. Since DoJoBa system can explore both the identity and the lexical information from the j-vector, it performs better than standard joint Bayesian systems.

Table 1: Performance of different systems on the evaluation set of RSR2015 part I in terms of equal error rate (EER %).

EER(%)	j-vector	joint Bayesian	jPLDA	DoJoBa
IW	0.95	0.02	0.02	0.02
TW	3.14	0.03	0.06	0.02
IC	7.86	3.61	3.12	2.97
Total	1.45	0.46	0.40	0.37

Table 2: Performance of different systems on the evaluation set of Huiting202 in terms of equal error rate (EER %).

EER(%)	j-vector	joint Bayesian	jPLDA	DoJoBa
IW	0.86	0.10	0.13	0.08
TW	6.71	0.04	0.07	0.04
IC	4.57	2.52	2.37	2.13
Total	1.37	0.45	0.36	0.31

5. Conclusions

In this paper we have proposed a double joint Bayesian (DoJoBa) analysis for j-vector verification. DoJoBa is related to joint Bayesian model, and can be thought of as a joint Bayesian analysis with multiple probability distributions attached to the features. The most important advantages of DoJoBa, compared to joint Bayesian, is that multiple information can be explicitly modeled and explored from the samples to improve verification performance; compared to jPLDA, DoJoBa can determine the latent dimension without tuning. Reported results showed that DoJoBa provided significant reduction in error rates over conventional systems in term of EER.

6. Acknowledgements

We would like to thank the four anonymous referees, whose suggestions improve the quality of this paper.

7. References

[1] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[2] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5115–5119.

[3] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, and Md Jahangir Alam, "JFA-based front ends for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1705–1709.

[4] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[6] Ehsan Variani, Xin Lei, Erik Mcdermott, and Ignacio Lopez Moreno, "Deep neural networks for small footprint text-dependent speaker verification," in *ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4052–4056.

[7] Hossein Zeinali, Hossein Sameti, Lukas Burget, Jan Cernocky, Nooshin Maghsoodi, and Pavel Matejka, "i-vector/HMM based text-dependent speaker verification system for RedDots challenge," in *INTERSPEECH*, 2016.

[8] Nanxin Chen, Yanmin Qian, and Kai Yu, "Multi-task learning for text-dependent speaker verification," in *INTERSPEECH*, 2015.

[9] Sergey Ioffe, "Probabilistic linear discriminant analysis," *Proc ECCV*, vol. 22, no. 4, pp. 531–542, 2006.

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

[11] Aleksandr Sizov, Kong Aik Lee, and Tomi Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," *S+SSPR 2014 Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition - Volume 8621*, pp. 464–475, 2014.

[12] Dong Chen, Xudong Cao, David Wipf, Fang Wen, and Jian Sun, "An efficient joint formulation for bayesian face verification," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 32–46, 2017.

[13] Yiyan Wang, Haotian Xu, and Zhijian Ou, "Joint bayesian gaussian discriminant analysis for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[14] Ziqiang Shi, Liu Liu, Mengjiao Wang, and Rujie Liu, "Multi-view (joint) probability linear discrimination analysis for j-vector based text dependent speaker verification," in *ASRU*, 2017.

[15] "Maximum likelihood estimation from incomplete data via the EM algorithm, author=Dempster, A. P., journal=Journal of the Royal Statistical Society, volume=39, number=1, pages=1-38, year=1977,," .

[16] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "Plda modeling in i-vector and supervector space for speaker verification," in *ACM International Conference on Multimedia, Singapore, November*, 2012, pp. 882–891.

[17] Yuan Liu, Yanmin Qian, Nanxin Chen, Tianfan Fu, Ya Zhang, and Kai Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[18] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks.," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

8. Appendix

This appendix provides the derivation of the formulae about the double joint Bayesian (DoJoBa) formulation.

8.1. Training of DoJoBa

Let $X = \{x_{ijk} : i = 1, ..., I; j = 1, ..., J; k = 1, ..., H_{ij}\}$. In order to find the parameters $\theta = \{\mu, \Sigma_u, \Sigma_v, \Sigma_\epsilon\}$ under which the data set X is most likely, the classical EM algorithm [15] is employed.

Let $X = \{x_{ijk} \in \mathbb{R}^D : i = 1, ..., I; j = 1, ..., J; k = 1, ..., H_{ij}\}$, $x_{ij} = \{x_{ijk} : k = 1, ..., H_{ij}\}$, and $x_i = \{x_{ijk} : j = 1, ..., J; k = 1, ..., H_{ij}\}$.

E steps: we need to calculate the expectations $\mathrm{E}_{U|X,\theta_t}[u_i]$, $\mathrm{E}_{V|X,\theta_t}[v_j]$, $\mathrm{E}_{U|X,\theta_t}[u_i u_i^T]$, $\mathrm{E}_{V|X,\theta_t}[v_j v_j^T]$, and $\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T]$.

For $\mathrm{E}_{U|X,\theta_t}[u_i]$ and $\mathrm{E}_{U|X,\theta_t}[u_i u_i^T]$, we have

$$
\begin{aligned}
& p(u_i|x_i, \theta) \\
\propto\ & p(x_i|u_i, \theta)p(u_i) = \left[\prod_{j=1}^{J}\prod_{k=1}^{H_{ij}} p(x_{ijk}|u_i, \theta)\right] p(u_i) \\
=\ & \left[\prod_{j=1}^{J}\prod_{k=1}^{H_{ij}} \mathcal{N}(x_{ijk}|\mu + u_i + v_i, \Sigma_\epsilon)\right] \mathcal{N}(u_i|0, \Sigma_u) \\
\propto\ & \{\prod_{j=1}^{J}\prod_{k=1}^{H_{ij}} \exp[-\frac{1}{2}(x_{ijk} - \mu - u_i - v_j)^T \Sigma_\epsilon^{-1} \\
\times\ & (x_{ijk} - \mu - u_i - v_j)]\} \exp\left(-\frac{1}{2}u_i^T \Sigma_u^{-1} u_i\right) \\
\propto\ & \exp\{\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}[(x_{ijk} - \mu)^T \Sigma_\epsilon^{-1}(u_i + v_j) \\
-\ & \frac{1}{2}(u_i + v_j)^T \Sigma_\epsilon^{-1}(u_i + v_j)] - \frac{1}{2}u_i^T \Sigma_u^{-1} u_i\} \\
\propto\ & \exp\{\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}(x_{ijk} - \mu - v_j)^T \Sigma_\epsilon^{-1} u_i \\
-\ & \frac{1}{2}u_i^T(\Sigma_u^{-1} + \Sigma_\epsilon^{-1}\sum_{j=1}^{J} H_{ij})u_i\},
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
& \mathrm{E}_{U|X,\theta_t}[u_i] \\
=\ & \left(\Sigma_u^{-1} + \Sigma_\epsilon^{-1}\sum_{j=1}^{J} H_{ij}\right)^{-1} \Sigma_\epsilon^{-1}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}(x_{ijk} - \mu - v_j).
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathrm{E}_{U|X,\theta_t}[u_i u_i^T] \\
=\ & \left(\Sigma_u^{-1} + \Sigma_\epsilon^{-1}\sum_{j=1}^{J} H_{ij}\right)^{-1} + \mathrm{E}_{U|X,\theta_t}[u_i]\mathrm{E}_{Z|X,\theta_t}[u_i]^T.
\end{aligned}
$$

Almost the same process for $\mathrm{E}_{V|X,\theta_t}[v_j]$ and $\mathrm{E}_{V|X,\theta_t}[v_j v_j^T]$.

For $\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T]$, indeed we have

$$
\begin{aligned}
& p(u_i, v_j | x_{ij}, \theta) \\
\propto\ & p(x_{ij}|u_i, v_j, \theta) p(u_i, v_j) = \left[\prod_{k=1}^{H_{ij}} p(x_{ijk}|u_i, v_j, \theta)\right] p(u_i, v_j) \\
=\ & \left[\prod_{k=1}^{H_{ij}} \mathcal{N}(x_{ijk}|\mu + u_i + v_i, \Sigma_\epsilon)\right] \mathcal{N}(u_i, v_j|0, \textbf{diag}[\Sigma_u, \Sigma_v]) \\
\propto\ & \{\prod_{k=1}^{H_{ij}} \exp[-\frac{1}{2}(x_{ijk} - \mu - u_i - v_j)^T \Sigma_\epsilon^{-1} \\
\times\ & (x_{ijk} - \mu - u_i - v_j)]\} \exp\left(-\frac{1}{2}u_i^T \Sigma_u^{-1} u_i - \frac{1}{2} v_j^T \Sigma_v^{-1} v_j\right) \\
\propto\ & \exp\{\sum_{k=1}^{H_{ij}} (x_{ijk} - \mu)^T \Sigma_\epsilon^{-1} (u_i + v_j) \\
-\ & \frac{1}{2}(u_i + v_j)^T \Sigma_\epsilon^{-1} (u_i + v_j) - \frac{1}{2} u_i^T \Sigma_u^{-1} u_i - \frac{1}{2} v_j^T \Sigma_v^{-1} v_j\} \\
\propto\ & \exp\{\sum_{k=1}^{H_{ij}} (x_{ijk} - \mu)^T \Sigma_\epsilon^{-1} \mathbf{B} z_{ij} \\
-\ & \frac{1}{2} z_{ij}^T (\textbf{diag}[\Sigma_u^{-1}, \Sigma_v^{-1}] + H_{ij} \mathbf{B}^T \Sigma_\epsilon^{-1} \mathbf{B}) z_{ij}\},
\end{aligned}
$$

where $z_{ij} = \begin{bmatrix} u_i \\ v_j \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix}$.

Thus we have

$$
\begin{aligned}
& \mathrm{E}_{Z|X,\theta_t}\left\{\begin{bmatrix} u_i \\ v_j \end{bmatrix}\right\} = \mathrm{E}_{Z|X,\theta_t}[z_{ij}] \\
=\ & \left(\textbf{diag}[\Sigma_u^{-1}, \Sigma_v^{-1}] + H_{ij} \mathbf{B}^T \Sigma_\epsilon^{-1} \mathbf{B}\right)^{-1} \mathbf{B}^T \Sigma_\epsilon^{-1} \sum_{k=1}^{H_{ij}} (x_{ijk} - \mu).
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathrm{E}_{Z|X,\theta_t}\left\{\begin{bmatrix} u_i u_i^T & u_i v_j^T \\ v_j u_i^T & v_j v_j^T \end{bmatrix}\right\} = \mathrm{E}_{Z|X,\theta_t}[z_{ij} z_{ij}^T] \\
=\ & \left(\textbf{diag}[\Sigma_u^{-1}, \Sigma_v^{-1}] + H_{ij} \mathbf{B}^T \Sigma_\epsilon^{-1} \mathbf{B}\right)^{-1} \\
+\ & \mathrm{E}_{Z|X,\theta_t}[z_{ij}] \mathrm{E}_{Z|X,\theta_t}[z_{ij}]^T.
\end{aligned}
$$

M steps: we update the values of the parameters $\theta = \{\mu, \Sigma_u, \Sigma_v, \Sigma_\epsilon\}$ and have

$$
\begin{aligned}
& \theta_{t+1} \\
:=\ & \arg\max_\theta Q(\theta|\theta_t) \\
=\ & \arg\max_\theta \mathrm{E}_{U,V|X,\theta_t}[\log \mathcal{L}(\theta; X, U, V)] \\
=\ & \arg\max_\theta \mathrm{E}_{U,V|X,\theta_t}[\log p(X, U, V|\theta)] \\
=\ & \arg\max_\theta \mathrm{E}_{U,V|X,\theta_t}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} \log p(x_{ijk}, u_i, v_j|\theta)\right] \\
=\ & \arg\max_\theta \mathrm{E}_{U,V|X,\theta_t}\{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} \log[p(x_{ijk}|u_i, v_j, \theta) \\
\times\ & p(u_i, v_j)]\} \\
=\ & \arg\max_\theta \mathrm{E}_{U,V|X,\theta_t}\{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} \\
& \log[\mathcal{N}(x_{ijk}|\mu + u_i + v_j, \Sigma_\epsilon)\mathcal{N}(u_i, v_j|0, \textbf{diag}[\Sigma_u, \Sigma_v])]\} \\
=\ & -\arg\min_\theta \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} \mathrm{E}_{U,V|X,\theta_t}\{\frac{1}{2}(\log|\Sigma_\epsilon| \\
+\ & \log|\Sigma_u| + \log|\Sigma_v|) \\
+\ & \frac{1}{2}(x_{ijk} - \mu - u_i - v_j)^T \Sigma_\epsilon^{-1} (x_{ijk} - \mu - u_i - v_j)) \\
+\ & \frac{1}{2} u_i^T \Sigma_u^{-1} u_i + \frac{1}{2} v_j^T \Sigma_v^{-1} v_j\} \\
=\ & \arg\max_\theta \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} [-\frac{1}{2}(\log|\Sigma_\epsilon| + \log|\Sigma_u| + \log|\Sigma_v|) \\
-\ & \frac{1}{2}(x_{ijk} - \mu)^T \Sigma_\epsilon^{-1} (x_{ijk} - \mu)] \\
+\ & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} (x_{ijk} - \mu)^T \Sigma_\epsilon^{-1} \left\{\mathrm{E}_{U|X,\theta_t}[u_i] + \mathrm{E}_{V|X,\theta_t}[v_j]\right\} \\
-\ & \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \{\mathrm{E}_{U|X,\theta_t}[u_i^T(\Sigma_\epsilon^{-1} + \Sigma_u^{-1})u_i] \\
+\ & 2\mathrm{E}_{U,V|X,\theta_t}[v_j^T \Sigma_\epsilon^{-1} u_i] + \mathrm{E}_{V|X,\theta_t}[v_j^T(\Sigma_\epsilon^{-1} + \Sigma_v^{-1})v_j]\}
\end{aligned}
$$

Take derivatives with respect to Σ_u^{-1}, Σ_v^{-1}, Σ_ϵ^{-1}, and μ and then equate these derivatives to zero to proved the update rules. The following is the detailed derivations.

For Σ_u^{-1}, we have

$$
\frac{\partial Q}{\partial \Sigma_u^{-1}} = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \left[\Sigma_u - \mathrm{E}_{U|X,\theta_t}[u_i u_i^T]\right]
$$

Setting $\frac{\partial Q}{\partial \Sigma_u^{-1}} = 0$, we have

$$
\Sigma_u = \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} 1} \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \mathrm{E}_{U|X,\theta_t}[u_i u_i^T].
$$

For Σ_v^{-1}, we have

$$
\frac{\partial Q}{\partial \Sigma_v^{-1}} = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \left[\Sigma_v - \mathrm{E}_{V|X,\theta_t}[v_j v_j^T]\right]
$$

Setting $\frac{\partial Q}{\partial \Sigma_v^{-1}} = 0$, we have

$$\Sigma_v = \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} 1} \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \mathrm{E}_{V|X,\theta_t}[v_j v_j^T].$$

For Σ_ϵ^{-1}, we have

$$\begin{aligned}
& \frac{\partial Q}{\partial \Sigma_\epsilon^{-1}} \\
= \quad & \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \left[\Sigma_\epsilon - (x_{ijk}-\mu)(x_{ijk}-\mu)^T\right] \\
+ \quad & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} (x_{ijk}-\mu)\left[\mathrm{E}_{U|X,\theta_t}[u_i]^T + \mathrm{E}_{V|X,\theta_t}[v_j]^T\right] \\
- \quad & \frac{1}{2}[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} (\mathrm{E}_{U|X,\theta_t}[u_i u_i^T] + 2\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T] \\
+ \quad & \mathrm{E}_{V|X,\theta_t}[v_j v_j^T])]
\end{aligned}$$

Setting $\frac{\partial Q}{\partial \Sigma_\epsilon^{-1}} = 0$ and rearranging, result in

$$\begin{aligned}
\Sigma_\epsilon \quad & = \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} 1} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \{(x_{ijk}-\mu)(x_{ijk}-\mu)^T \\
- \quad & 2(x_{ijk}-\mu)[\mathrm{E}_{U|X,\theta_t}[u_i]^T + \mathrm{E}_{V|X,\theta_t}[v_i]^T] \\
+ \quad & \Big(\mathrm{E}_{U|X,\theta_t}[u_i u_i^T] + 2\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T] + \mathrm{E}_{V|X,\theta_t}[v_j v_j^T]\Big)\}.
\end{aligned}$$

For μ, we have

$$\begin{aligned}
& \frac{\partial Q}{\partial \mu} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} -(x_{ijk}-\mu)^T \Sigma_\epsilon^{-1} \\
- \quad & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \left[\mathrm{E}_{U|X,\theta_t}[u_i]^T + \mathrm{E}_{V|X,\theta_t}[v_i]^T\right]\Sigma_\epsilon^{-1}
\end{aligned}$$

Setting $\frac{\partial Q}{\partial \mu} = 0$, we have

$$\begin{aligned}
& \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} -(x_{ijk}-\mu)^T \\
- \quad & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} \left[\mathrm{E}_{U|X,\theta_t}[u_i]^T + \mathrm{E}_{V|X,\theta_t}[v_j]^T\right] = 0
\end{aligned}$$

Since $\mathrm{E}_{U|X,\theta_t}[u_i] \approx 0$ and $\mathrm{E}_{V|X,\theta_t}[v_j] \approx 0$, we have $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} -(x_{ij}-\mu) = 0$, that is

$$\mu = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} x_{ijk}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ijk}} 1}.$$