Interspeech 2018
2-6 September 2018, Hyderabad

Double Joint Bayesian Modeling of DNN Local I-Vector for Text Dependent Speaker Verification with Random Digit Strings

Ziqiang Shi, Huibin Lin, Liu Liu, Rujie Liu

Fujitsu Research and Development Center
shiziqiang@cn.fujitsu.com

Abstract

Double joint Bayesian is a recently introduced analysis method that models and explores multiple information explicitly from the samples to improve the verification performance. It was recently applied to voice pass phrase verification, result in better results on text dependent speaker verification task. However little is known about its effectiveness in other challenging situations such as speaker verification for short, text-constrained test utterances, e.g. random digit strings. Contrary to conventional joint Bayesian method that cannot make full use of multi-view information, double joint Bayesian can incorporate both intra-speaker/digit and inter-speaker/digit variation, and calculated the likelihood to describe whether the features having all labels consistent or not. We show that double joint Bayesian outperforms conventional method on modeling DNN local (digit-dependent) i-vectors for speaker verification with random prompted digit strings. Since the strength of both double joint Bayesian and conventional DNN local i-vector appear complementary, the combination significantly outperforms either of its components.

Index Terms: speaker verification, joint Bayesian analysis, DNN i-vector

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, 3]. Text dependent speaker verification has wide applications in many areas, including smart human-machine interface, security, forensic, telephone banking, and so on.

Typical text-dependent speaker verification uses fixed phrase for each user and hence, enrollment and test phrases are matched. For this scenario it is possible that utterance from a user can be recorded beforehand by an imposter and then play it back. This spoofing or attack can be avoided to a certain extent by sharing the same phonetic content but with different context between training and test utterances, for example the user is prompted to utter a digit strings randomly chosen by the system. In this anti-spoofing scenario, the speaker is usually prompted to utter all of 10 digits several times during enrollment and test utterances contain a subset of the digits. This work is tested on part III of the RSR2015 database [2] which is designed to evaluate the ability of a system to deal with this kind of scenario.

Previous methods regarding speaker verification with random prompted digit string can be grouped into two categories. The first category is based on the traditional state-of-the-art Gaussian mixture model represented universal background model (GMM-UBM) and joint factor analysis (JFA) approach: Larcher et al. [2] use a Hidden Markov Model (HMM) system termed HiLAM to model each speaker and each state corresponding one of the 10 digits; Stafylakis et al. [4] propose to use JFA to extract the global utterance vector and local digit vector, which are fed into a joint density backend.

In the second category, deep models are ported to speaker verification: deep neural network (DNN) is used to estimate the frame posterior probabilities [5]; DNN as a feature extractor for the utterance level representation [6]; Matejka et al. [7] have shown that using bottleneck DNN features (BN) concatenated to other acoustic features outperformed the DNN method for text-dependent speaker verification; end-to-end deep learning jointly optimizes the speaker representations and models [3]; multi-task deep learning jointly learns both speaker identity and text information [8].

This paper is based on the work of Lei et al. [5], in which a DNN is trained for phone recognition instead of a GMM-UBM to produce frame posteriors for the computation of i-vectors [9] and the work of [10, 11], in which the state-of-the-art joint Bayesian and probability linear discriminant analysis (PLDA) [12, 13] approach is extended to model the two voice pass phrases jointly with an appropriate prior that considers intra-speaker/phrase and inter-speaker/phrase variation over the speaker pairs and phrase pairs at the same time.

Since the digit vocabulary of RSR2015 part III is fixed small and limited, and further more that different speakers have different pronunciations of the same digits, thus in this work we propose to extract local digit-dependent DNN i-vectors for speaker verification. Such local DNN i-vectors potentially have different kinds of labels including a speaker latent variable and a digit latent variable. This means the two latent variables related to speaker and local digit pronunciations have equal importance, and both variables are tied across all samples sharing a certain label. Double Joint Bayesian (DoJoBa) [10] is employed to do the verification of such DNN local i-vectors. The relationship between DoJoBa and standard joint Bayesian is analogous to that between joint factor analysis and factor analysis.

Our contribution is two-fold. Firstly we propose to use DNN i-vector framework [5] to extract digit dependent DNN i-vectors, which outperforms the utterance level features. Secondly we propose to use DoJoBa to explicitly and jointly model the multi-view information from digit samples, such as certain individual saying certain digit, which leads to a significant improvement for the speaker verification performance in RSR2015 part III.

The remainder of this paper is organized as follows: Section 2 describes the DNN local i-vector/DoJoBa approach; The detail experimental results and comparisons are presented in Section 3 and the whole work is summarized in Section 4.

10.21437/Interspeech.2018-1103

2. Model description

In this section we first review the DNN local i-vector representation of utterance, and then present DoJoBa for the modeling and likelihood ratio scoring of the DNN local i-vectors.

2.1. DNN local i-vector

Classical i-vector approach [9] assumes that the t-th frame feature from the i-th utterance is generated by the following process

$$x_{t,i} \sim \sum_k \pi_{t,i,k} \mathcal{N}(\mu_k + T_k w_i, \Sigma_k) \tag{1}$$

where $\mathcal{N}(\mu, \Sigma)$ represents a Gaussian with mean μ and covariance Σ, T_k and the latent vector w_i are the total variability subspace and i-vector respectively, $\pi_{t,i,k}$ is the posterior of $x_{t,i}$ be generated by the k-th component. In order to train the T_k and to extract the i-vector w_i the following necessary and sufficient zero-order, first-order and second-order statistics

$$\begin{aligned} N_{k,i} &= \sum_t \gamma_{k,t,i} x_{t,i} \\ F_{k,i} &= \sum_t \gamma_{k,t,i} x_{t,i} \\ S_{k,i} &= \sum_t \gamma_{k,t,i} x_{t,i} x_{t,i}^T \end{aligned} \tag{2}$$

needed to be computed respectively, where $\gamma_{k,t,i}$ is the posterior of $x_{t,i}$ with respect to the k-th Gaussian.

In [5], Lei et al. proposed to use the DNN to replace the GMM to compute the posterior $\gamma_{k,t,i}$ of the frames with respect to each of the classes in the model. While in the case of the GMM in traditional i-vector extractor, the classes are the individual Gaussian from a mixture model, in the case of the DNN the classes are senones. Given an utterance, the statistics in (2) now can be updated by using the new posterior probabilities $\gamma_{k,t,i}$ of the senone classes. Then these sufficient statistics are used to train the total variability matrix T_k and the i-vector w_i, which is called the DNN i-vector.

In the task of speaker verification with random prompted digit strings since different speakers have own characterizations in pronouncing each same digit. This digit-dependent characterization will definitely help in this task. In order to extract DNN i-vectors for each digits (will be called DNN local i-vector in this work), we trained a DNN-HMM automatic speech recognition (ASR) system based on the Librispeech corpus [14] for two purposes: one is to do segmentation of the utterances in to digits (more specifically to do the alignment between the utterance and the prompt digit string); and the other is to train a DNN used in the extraction of the DNN local i-vector.

2.2. Double Joint Bayesian

We assume that the training data is obtained from I speakers saying J digits each with H_{ij} sessions. We denote the DNN local i-vector of the k'th session of the i'th speaker saying j'th digit by x_{ijk} (please do not confuse with the notation x in 1, here x_{ijk} is indeed the w in 1). We model the digit dependent feature generation by the process [10]:

$$x_{ijk} = \mu + u_i + v_j + \epsilon_{ijk}. \tag{3}$$

The model comprises two parts: 1, the signal component $\mu + u_i + v_j$ which depends only on the speaker and digit, rather than on the particular DNN local i-vector (i.e. there is no dependence on k); 2, the noise component ϵ_{ijk} which is different for every DNN local i-vector of the speaker/digit and represents within-speaker/digit 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 digit 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 [15] is analogous (not exactly) to that between joint PLDA [11] 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 [17] is employed.

2.2.1. *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

$$\begin{aligned} &\mathrm{E}_{U|X,\theta_t}[u_i] = \\ &\left(\Sigma_u^{-1} + \Sigma_\epsilon^{-1} \sum_{j=1}^{J} H_{ij} \right)^{-1} \Sigma_\epsilon^{-1} \sum_{j=1}^{J} \sum_{k=1}^{H_{ij}} (x_{ijk} - \mu - v_j). \end{aligned} \tag{4}$$

and

$$\begin{aligned} &\mathrm{E}_{U|X,\theta_t}[u_i u_i^T] = \\ &\left(\Sigma_u^{-1} + \Sigma_\epsilon^{-1} \sum_{j=1}^{J} H_{ij} \right)^{-1} + \mathrm{E}_{U|X,\theta_t}[u_i] \mathrm{E}_{U|X,\theta_t}[u_i]^T. \end{aligned} \tag{5}$$

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

$$\mathrm{E}_{U,V|X,\theta_t}\left\{\begin{bmatrix}u_i u_i^T & u_i v_j^T\\ v_j u_i^T & v_j v_j^T\end{bmatrix}\right\} = \tag{6}$$

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

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

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

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

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

$$\Sigma_\epsilon = \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}} 1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{H_{ij}}\{(x_{ij} - \mu)(x_{ijk} - \mu)^T$$

$$-2(x_{ijk} - \mu)[\mathrm{E}_{U|X,\theta_t}[u_i]^T + \mathrm{E}_{V|X,\theta_t}[v_i]^T]$$

$$+\mathrm{E}_{U|X,\theta_t}[u_i u_i^T] + 2\mathrm{E}_{U,V|X,\theta_t}[u_i v_j^T] + \mathrm{E}_{V|X,\theta_t}[v_j v_j^T]\},$$

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 (4), (5) and (6).

2.2.2. *Likelihood Ratio Scores*

We treat the verification as a kind of hypothesis testing problem with the null hypothesis $\mathcal{H}_0$ where two DNN local i-vectors have the same speaker and digit 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 $\mathcal{M}_3$). Given a test DNN local i-vector x_t and an enrolled DNN local i-vector x_s, and let the priori probability of the models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ as $p_1 = P(\mathcal{M}_1|\mathcal{H}_1)$, $p_2 = P(\mathcal{M}_2|\mathcal{H}_1)$, $p_3 = P(\mathcal{M}_3|\mathcal{H}_1)$, then the likelihood ratio score is

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

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

where

$$\mathbf{X} = P(x_t, x_s|\mathcal{H}_1) = P(x_t, x_s|\mathcal{M}_1)P(\mathcal{M}_1|\mathcal{H}_1)$$

$$+P(x_t, x_s|\mathcal{M}_2)P(\mathcal{M}_2|\mathcal{H}_1) + P(x_t, x_s|\mathcal{M}_3)P(\mathcal{M}_3|\mathcal{H}_1)$$

$$= p_1\mathcal{N}(\begin{bmatrix}x_t\\ x_s\end{bmatrix} \mid \begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{bmatrix}\Sigma_u + \Sigma_v + \Sigma_\epsilon & \Sigma_v\\ \Sigma_v & \Sigma_u + \Sigma_v + \Sigma_\epsilon\end{bmatrix})$$

$$+p_2\mathcal{N}(\begin{bmatrix}x_t\\ x_s\end{bmatrix} \mid \begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{bmatrix}\Sigma_u + \Sigma_v + \Sigma_\epsilon & \Sigma_u\\ \Sigma_u & \Sigma_u + \Sigma_v + \Sigma_\epsilon\end{bmatrix})$$

$$+p_3\mathcal{N}(x_t|\mu, \Sigma_u + \Sigma_v + \Sigma_\epsilon)\mathcal{N}(x_s|\mu, \Sigma_u + \Sigma_v + \Sigma_\epsilon).$$

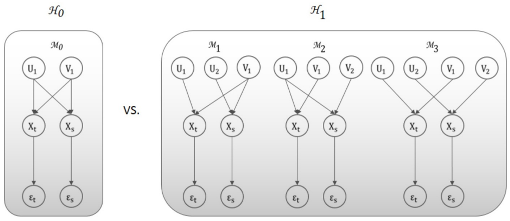

Figure 1: *Verification by comparing the likelihood of the data under different hypotheses. Under the null hypothesis $\mathcal{H}_1$, the feature x_t and x_s do not match. Under the hypothesis $\mathcal{H}_0$ they match.*

Notice that like standard joint Bayesian model [15], we do not calculate a point estimate of hidden variable. Instead we compute the probability that the two multi-label vectors had the same hidden variables, regardless of what this actual latent variable was. It is also to note that since on RSR2015 each test utterance contains 5 digits, thus each trial contains 5 pairs of DNN local i-vector during testing and results in 5 digits log likelihood ratios, which will be simply averaged to obtain the overall log likelihood ratio of the trial.

2.3. Score normalization

In order to transform log likelihood ratio scores from different speakers into a similar range by using

$$s' = \frac{s - \mu_I}{\sigma_I}$$

so that a common threshold can be used, where μ_I and σ_I are the approximated mean and standard deviation of the impostor score distribution respectively. We tried three score normalization method: zero normalization (z-norm) uses a batch of non-target utterances against the target model to compute the mean μ_I and standard deviation σ_I; test normalization (t-norm) uses the unknown speaker's feature vectors against a set of impostor models to compute the statistics; the zero and test normalized scores are finally averaged to form the s-normalized scores [4].

3. Evaluation and discussion

In this section, we describe the experimental setup and results for the proposed method on the public RSR2015 part III English corpus [2].

3.1. Experimental setup

RSR2015 corpus [2] was released by I2R, and it is used to evaluate the performance of different speaker verification systems. In this work, we follow the setup of [4], the part

III of RSR2015 is used for the testing of our method. Part III of RSR2015 contains 300 speakers speaking in English and chosen so that they form a representative sample of the Singaporean population. All speech files are of 16kHz. The gender distribution is balanced on the data set (157 male and 143 female). Six mobile devices were used for the recordings that took place under a typical office environment. The speakers are divided into three disjoint groups refereed to as background, development and evaluation, of 97, 97 and 106 speakers respectively. Each speaker model is enrolled with 3 10-digit utterances, recorded with the same handset, while each speaker contributes 3 different speaker models. Each test utterance contains a quasi-random string of 5 digits, one out of 52 unique strings. For both types of utterances, the digit string is given and the verification algorithm may use it. In Table I, the number of trials used for the experiments are given for each set and gender[1].

Table 1: *Trial statistics for RSR2015 digits per set and gender.*

Set	Gender	#target	#nontarget
Dev	Male	5154	251310
Dev	Female	5052	231155
Eval	Male	5943	332863
Eval	Female	5283	253584

The input feature is 39-dimensional Mel-frequency cepstral coefficients (MFCC, 13 static including the log energy + 13 Δ + 13 $\Delta\Delta$) are extracted and normalized using utterance-level mean and variance normalization. Then the frame-senon pairs aligned by the GMM-HMM system will be used to train a fully connected DNN. The DNN has 6 hidden layers (with sigmoid activation function) of 2048 nodes each. The output layer, which is the classification layer, is a softmax of dimension 9020 i.e., the output layer computes posteriors for 9020 triphone tied states (senones).

3.2. Results and discussion

Three systems are evaluated and compared across above conditions:

- **DNN i-vector**: the standard DNN i-vector system with PLDA [18].
- **DNN local i-vector**: the DNN local i-vector system with average log likelihood scores from the DoJoBa system across the digits in the utterance.
- **DoJoBa**: the DNN local i-vector with the DoJoBa system described in Section 2.

Table 2: *Performance of different systems on the development set of RSR2015 part III in terms of equal error rate (EER %).*

EER(%),m/f	DNN i-vector	DNN local iv	DoJoBa
w/o norm	3.26/3.48	2.75/2.92	2.19/2.33
z-norm	2.54/2.86	2.38/2.74	2.07/2.28
t-norm D	2.39/2.57	2.23/2.55	1.94/2.19
s-norm	2.36/2.66	2.32/2.43	1.88/2.05

Table 3: *Performance of different systems on the evaluation set of RSR2015 part III in terms of equal error rate (EER %).*

EER(%),m/f	DNN i-vector	DNN local iv	DoJoBa
w/o norm	2.90/3.08	2.38/2.53	1.84/1.93
z-norm	2.23/2.55	2.05/2.37	1.69/1.91
t-norm	2.03/2.51	1.85/2.19	1.62/1.79
s-norm	1.99/2.35	1.93/2.12	1.48/1.66

Table 2 and Table 3 compare the performances of all above-mentioned systems in terms of equal error rate (EER) on the development and evaluation sets of RSR2015 part III respectively. Obviously digit dependent DNN local i-vector is superior to the standard DNN i-vector in this task, regardless of the test database and the backend when compared with results in [18].

Since DoJoBa system can explore both the identity and the digit information from the DNN local i-vector, it constantly performs better than standard DNN i-vector and joint Bayesian systems. It can be seen from the results that the DNN local i-vector with the DoJoBa system can obtain the state-of-the-art performance.

4. Conclusion

In this paper we investigated the effectiveness of a double joint Bayesian modeling for DNN local i-vector on the task of speaker verification with random prompted digit strings. By explicit modeling and exploring the difference in pronouncing of each digit by different speakers the new framework outperformed the DNN i-vector/PLDA approach.

5. References

[1] A. L. Higgins, L. G. Bahler, and J. E. Porter, “Speaker verification using randomized phrase prompting,” *Digital Signal Processing*, vol. 1, no. 2, pp. 89–106, 1991.

[2] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and rsr2015,” *Speech Communication*, vol. 60, pp. 56–77, 2014.

[3] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5115–5119.

[4] T. Stafylakis, M. J. Alam, and P. Kenny, “Text-dependent speaker recognition with random digit strings,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1194–1203, 2016.

[5] 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), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[6] E. Variani, X. Lei, E. Mcdermott, and I. L. 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] H. Zeinali, H. Sameti, L. Burget, J. Cernocky, N. Maghsoodi, and P. Matejka, “i-vector/hmm based text-dependent speaker verification system for reddots challenge,” in *INTERSPEECH*, 2016.

[8] N. Chen, Y. Qian, and K. Yu, “Multi-task learning for text-dependent speaker verificaion,” in *INTERSPEECH*, 2015.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,”

[1] The numbers of trials are the same as the work [2] and a little different from [4] of Dr. Stafylakis, since they rejected some utterances due to duration and SNR constrains.

IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[10] Z. Shi, M. Wang, L. Liu, H. Lin, and R. Liu, "A double joint bayesian approach for j-vector based text-dependent speaker verification." *arXiv preprint:1711.06434*, 2017.

[11] Z. Shi, L. Liu, M. Wang, and R. Liu, "Multi-view (joint) probability linear discrimination analysis for j-vector based text dependent speaker verification," in *ASRU*, 2017.

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

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

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[15] D. Chen, X. Cao, D. Wipf, F. Wen, and J. 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.

[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] A. P. Dempster, "Maximum likelihood estimation from incomplete data via the em algorithm (with discussion," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[18] J. Zhong, W. Hu, F. K. Soong, and H. Meng, "Dnn i-vector speaker verification with short, text-constrained test utterances." in *Interspeech 2017*, 2017, pp. 1507–1511.