ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Localized discriminative Gaussian process latent variable model for text-dependent speaker verification

Nooshin Maghsoodi[1], Hossein Sameti[1] and Hossein Zeinali[1,2]

1- Sharif University of Technology - Dept of Computer Engineering Tehran - Iran
2- Brno University of Technology - Dept of Information Technology Brno - Czech Republic

Abstract. The duration of utterances is one of the effective factors on the performance of speaker verification systems. Text dependent speaker verification suffers from both short duration and unmatched content between enrollment and test segments. In this paper, we use Discriminative Gaussian Process Latent Variable Model (DGPLVM) to deal with the uncertainty caused by short duration. This is the first attempt to utilize Gaussian Process for speaker verification. Also, to manage the unmatched content between enrollment and test segments we proposed the localized-DGPLVM that trains DGPLVM for each phrase in dataset. Experiments show the relative improvement of 27.4% in EER on RSR2015.

1 Introduction

Speaker verification is the process of acceptance or rejection of a claim of identity by comparing the speaker models generated from the enrollment and test utterances. If the lexicon used for the test utterances is a subset of the phrases used in enrollment step, the task is called text dependent [1]. Otherwise, the process can be defined as text independent. In the case of exact content match between enrollment and test utterances in text dependent speaker verification we achieve higher accuracy than text independent verification. However, if there is a mismatch between test and enrollment, the performance will drastically degrade because of short duration in text dependent task. Recently, the sate-of-the-art method in text independent speaker verification area is the i-vector [2] framework that when is used with probabilistic linear discriminant analysis (PLDA) for session variability compensation results in considerable improvements [3]. However, the success of this paradigm in text dependent verification is questionable. In fact, short length utterances in text dependent speaker verification cause i-vectors to tend to zero [4]. The reason is that the i-vector extraction method due to using MAP point estimate for i-vector calculation ignores the posterior covariance. This covariance matrix that could be a representation of the estimation uncertainty is a function of the inverse of the zero order statistics, the number of frames that are aligned to the UBM components. So, in short duration speaker verification the covariance will be greater and its ignorance is not reasonable.

One of the successful attempts to deal with this problem is propagating the uncertainty to the backend of the system. In [5-7], authors propose a modified version of PLDA that integrates the uncertainty of i-vector estimation into the PLDA model. Kenny et al. investigated phrase-dependent version of PLDA with uncertainty propagation in [8] to adapt their method to text dependent verification. Cumani et al. in [9] have proposed using i-vector full posterior distribution in PLDA model instead of its point estimation and derived the likelihood based on this posterior. In more recent studies [10-13], Kenny et al. used Joint Factor Analysis (JFA) based front end and to handle the uncertainty examined variational method to estimate the hidden variables. They did their experiments both in low dimension and supervector dimensional space. Experimental testbed in [10, 11] was RSR2015 part III [1] and to overcome the unmatched content problem the authors used tied mixture model to segment utterances into digits. In [11], the features passed to joint density backend as an analogue for PLDA and in [10], to consider the uncertainty of point estimates an i-vector based backend was proposed. A new scheme based on using i-vectors in text-prompted speaker verification is presented in [14]. It trains i-vector extractor and UBM for each word in dataset lexicon (i.e. Persian month names) separately.

In this paper, we propose a method based on Discriminative Gaussian Process Latent Variable Model (DGPLVM) [15] to learn a nonlinear mapping between the supervector space obtained from Universal Background Model (UBM) and a low dimensional latent space. The rationale behind this approach is that the large lexical variation in short duration utterances can be encoded by Gaussian process covariance function. On the other hand, the inter-speaker variability can be compensated using a discriminative prior for the latent space. In [16], authors present a principled multi-task learning approach based on DGPLVM for face verification. Also, GPLVM with a shared discriminative prior is proposed by Eleftheriadis et al. for multi-view facial expression recognition [17-19]. In fact, using Gaussian process for speaker verification is introduced in this paper for the first time and our novelty is using a set of phrase-localized Gaussian processes.

The rest of this paper is structured as follows. In Section 2 we give a short overview of DGPLVM. Section 3 details proposed speaker verification based on localized-DGPLVM. The description of the dataset and experimental results are given in Section 4. Finally, Section 5 provides a brief conclusion of this paper.

2 Discriminative Gaussian process latent variable model

Gaussian process is a stochastic process that finds a distribution over a set of functions. In fact, it is a generalization of the Gaussian distribution that is completely defined by a mean and covariance function [20].

The Gaussian Process Latent Variable Model (GPLVM) can be considered as a mapping between high dimensional input data and a low dimensional manifold. Let $X = \{x_1, x_2, \dots, x_N\}$, where $x_i \in \mathbb{R}^D$, be the input data and $Z = \{z_1, z_2, \dots, z_N\}$be the representation of data in the latent space, $z_i \in \mathbb{R}^d$ $(d \ll D)$. The GPLVM to find the locale of data in the latent space maximizes the posterior probability of the latent variable given the observation and the parameters, θ.To compute the posterior probability, we should compute the likelihood at first. It can be written as:

$$p(X|Z,\theta) = \frac{1}{\sqrt{(2\pi)^{ND}|\boldsymbol{K_X}|^D}} \exp(-\frac{1}{2} tr(\boldsymbol{K_X}^{-1} X X^T)),$$

where the elements of $\mathbf{K_X}$, the covariance function, is calculated based on a kernel function as $(\boldsymbol{K_X})_{i,j} = k_X(z_i, z_j)$. Considering the prior of latent variable as a zero mean, unit covariance Gaussian distribution, the logarithm of the posterior will be:

$$\mathcal{L} = \mathcal{L}_r + \sum_i \theta_i + \sum_i \frac{1}{2}||z_i||^2,$$
$$\mathcal{L}_r = \frac{D}{2}\ln|\boldsymbol{K_X}| + \frac{1}{2}tr\left(\boldsymbol{K_X}^{-1}XX^T\right).$$

By varying the prior probability over latent variable, different versions of GPLVM can be resulted. A category of these prior probabilities is discriminative priors. Since the GPLVM with spherical Gaussian prior is unsupervised, to integrate the label information of different classes, LDA based prior was proposed in [15]. LDA is a technique to transfer features to a new space by finding a set of discriminant axes. The objective function in this method minimizes inter-class variability while maximizes between-class variability in the new feature space. Denoting S_b and S_w between-class and within-class scatter matrices respectively, we have:

$$J(Z) = tr(S_w^{-1}S_b),$$
$$S_b = \sum_{i=1}^{S}(\bar{z}_i - \bar{z})(\bar{z}_i - \bar{z})^T,$$
$$S_w = \sum_{i=1}^{S}\sum_{s=1}^{n_s}(z_i^s - \bar{z}_i)(z_i^s - \bar{z}_i)^T,$$

where $\bar{z}_i$ is the mean of the data belonging to class i and $\bar{z}$ is the mean of all samples. Also, S is the class number and n_s indicates the number of all the data samples in class i. Changing the prior in GPLVM to be based on this objective function, we obtain:

$$p(Z) = \frac{1}{Z_b}\exp(-\frac{1}{\sigma^2}J^{-1}),$$

where Z_b represents normalization constant and σ is a global scaling of the prior.

Using this prior, the data points from the same class will be close in the latent space while the data from different classes will be far from each other. We will see in the next section that this kind of prior can help us to compensate the session variability in speaker verification.

3 Proposed speaker verification system based on DGPLVM

In order to learn low dimensional discriminative features from the speaker supervector space, we proposed using DGPLVM instead of factor analysis based approaches (e. g. i-vector and JFA). We believe that this model can improve the performance of speaker verification system when the utterances are too short. We also proposed to train DGPLVMs that are localized to single digits in our dataset.

Training separate subsystems containing UBM and i-vector extractor for each of the phrases in a Persian text-prompted dataset was first introduced in [14]. RSR2015 digit part is a text-prompted dataset which contains utterances with random sequence of digits. So, the vocabulary includes 10 digits. In this paper, we train a UBM for each digit in the dataset like [14] and due to considering the temporal order of speech frames that provides useful information in text dependent speaker verification, HMM is used as UBM. Then, for each utterance the first order statistics, the supervectors obtained from concatenating the centralized mean of the frames aligned to each GMM component of HMM states, is extracted. Therefore, the UBM output will be a set of supervectors per digit. Subsequently, each set of supervectors is given to a DGPLVM to train it. At the evaluation step, after segmenting the test utterances to digits, dimension reduction will be done using the Gaussian process corresponding to each of the segmented digits. Finally, the score is computed as a linear combination of the scores corresponding to the uttered digits in the test utterance. The scores are normalized by t-norm before combination. In this way, by using separate digit-level GPLVM the content variability is compensated and discriminative prior compensates the session variability. Figure 1 shows the block diagram of the different steps in the proposed method.

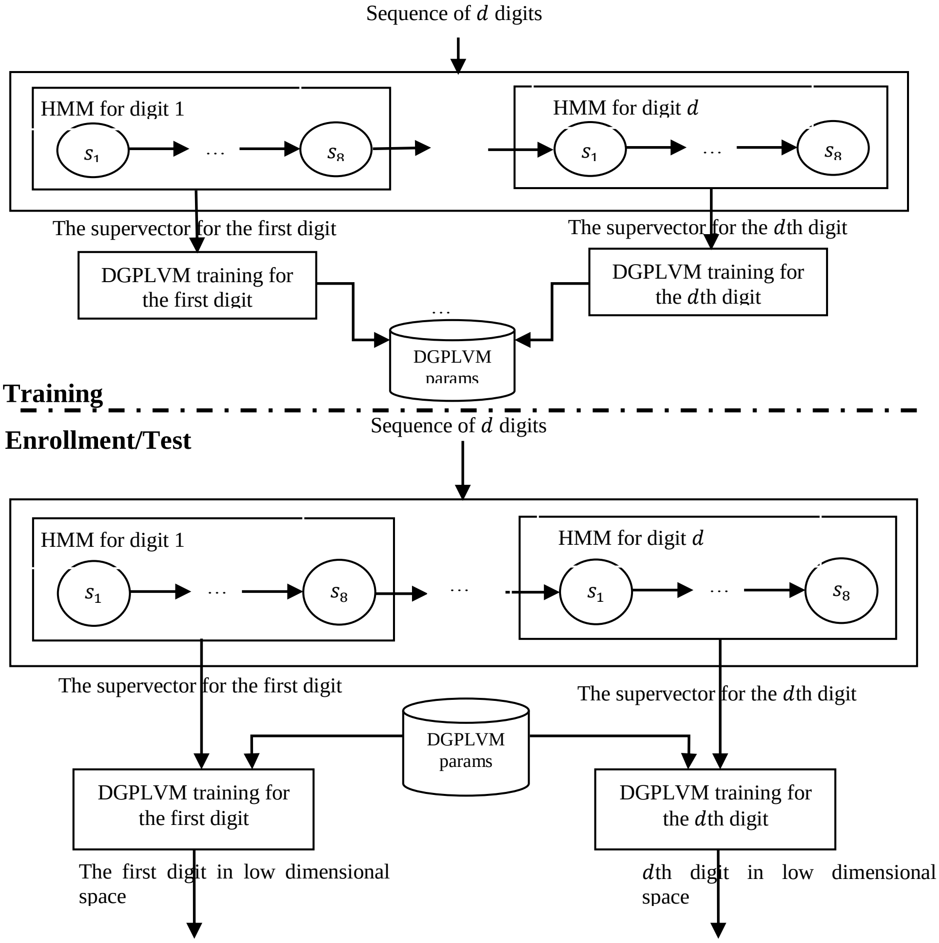
Sequence of d digits
HMM for digit 1
s_1
s_8
HMM for digit d
The supervector for the first digit
The supervector for the dth digit
DGPLVM training for the first digit
DGPLVM training for the dth digit
DGPLVM params
Training
Enrollment/Test
Sequence of d digits
HMM for digit 1
HMM for digit d
The supervector for the first digit
The supervector for the dth digit
DGPLVM params
DGPLVM training for the first digit
DGPLVM training for the dth digit
The first digit in low dimensional space
dth digit in low dimensional space

Fig. 1: Block diagram for the proposed system

4 Experiments

In the following the experimental results obtained from the implementation of the proposed method in this paper are described.

4.1 Experimental set-up and dataset

In the third part of RSR2015 all the speakers (i.e. 300 speakers) uttered 3 random sequences of 10 digits and 10 random sequences of five digits. The 10-digit sequences from evaluation set recorded in three sessions are used for model training and the test set contains all the 5-digit sequences. The experiments in this paper are limited to only female part of this dataset and the model is trained on the background set (*bkg*).

The experiments were based on feature vectors of length 60 including 19 PLP features together with log energy and their first and second derivatives. The window length and frame shift were 25 and 15 ms respectively. The features are extracted using HTK and CMVN was applied. HMMs with 8 states and 8 components were used as UBM and for digit level segmentation. Latent space dimensionality was set to 300 and the cosine distance based scores were normalized using gender dependent t-norm.

In this paper, Performance is represented in terms of equal error rate (EER) and minimum decision cost function (DCF). The parameters of DCF have those values used for NIST 2008 speaker recognition evaluation. To implement some parts of our verification system we used the MSR open source toolbox [21] and DS-GPLVM source code [17-19].

4.2 Results

The results of our method are compared with the state-of-the-art JFA-based methods on RSR2015 [10, 11] and the results of the baseline GMM-UBM method reported in [11]. The summary of the best results reported in [10, 11] are represented in Table 1. This table also summarizes the results obtained from the baseline system and our proposed method using discriminative prior and without it.

It is apparent from Table 1 that the localized-GPLVM with spherical Gaussian prior outperforms the GMM_UBM system and when we use the discriminative prior the system exhibits better performance than the state-of-the-art methods [10, 11]. Indeed, by using our localized-DGPLVM scheme 27.4% relative improvement in EER as well as 5.1% relative improvement in DCF over the best result from JFA-based method are observed.

System	EER (%)	Min DCF
digit dependent i-vector based backend [10]	5.9	0.297
digit dependent z-vector with fusion [11]	6.08	0.291
Baseline GMM-UBM [11]	8.36	0.383
Localized-GPLVM	7.49	0.363
Localized-GPLVM with discriminative prior	**4.28**	**0.276**

Table 1: Summary of experimental results for female part of RSR2015 digits, *eval* set.

5 Conclusion

RSR2015 digits dataset has two constraints, first, constraint of content variability between test and enrollment utterances and second, very short speech segments. In this paper, to overcome the problem of uncertainty causing by short duration, we investigated using GPLVM with discriminative prior. The discriminative prior helps us to find a subspace that compensates session variability. Also, to handle the unmatched content, we have trained all parts of the verification system in digit level.

In fact, in our system we have separate Gaussian process based models which are trained locally per digit. The experimental results showed that this method improves the state-of-the-art speaker verification system on RSR2015 digits part.

References

[1] 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.

[2] 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, pp. 788-798, 2011.

[3] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey*, 2010.

[4] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "I-Vector/PLDA Variants for Text-Dependent Speaker Recognition," ed: preparation, 2013.

[5] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 7649-7653, 2013.

[6] W. Cai, M. Li, L. L. Li, and Q. Hong, "Duration Dependent Covariance Regularization in PLDA Modeling for Speaker Verification," in *INTERSPEECH*, 2015.

[7] Q. Hong, L. L. Li, M. Li, L. Huang, L. Wan, and J. Zhang, "Modified-prior PLDA and Score Calibration for Duration Mismatch Compensation in Speaker Recognition System," in *INTERSPEECH*, 2015.

[8] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using PLDA with uncertainty propagation," *in INTERSPEECH*, pp. 3684-3688, 2013.

[9] S. Cumani, O. Plchot, and P. Laface, "On the use of i–vector posterior distributions in Probabilistic Linear Discriminant Analysis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, pp. 846-857, 2014.

[10] P. Kenny, T. Stafylakis, M. J. Alam, and M. Kockmann, "An I-Vector Backend for Speaker Verification," in *INTERSPEECH*, 2015.

[11] T. Stafylakis, P. Kenny, J. Alam, and M. Kockmann, "JFA for Speaker Recognition with Random Digit Strings," in *INTERSPEECH*, 2015.

[12] P. Kenny, T. Stafylakis, P. Ouellet, and M. J. Alam, "JFA-based front ends for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 1705-1709, 2014.

[13] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, and M. Kockmann, "Joint Factor Analysis for Text-Dependent Speaker Verification," *Odyssey*, 2014.

[14] H. Zeinali, E. Kalantari, H. Sameti, and H. Hadian, "Telephony text-prompted speaker verification using i-vector representation," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 4839-4843, 2015.

[15] R. Urtasun and T. Darrell, "Discriminative Gaussian process latent variable model for classification," in *Proceedings of the 24th international conference on Machine learning*, pp. 927-934, 2007.

[16] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with GaussianFace," *arXiv preprint arXiv:1404.3840*, 2014.

[17] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Shared gaussian process latent variable model for multi-view facial expression recognition," in *Advances in Visual Computing*, ed: Springer, pp. 527-538, 2013.

[18]S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative Shared Gaussian Processes for Multiview and View-Invariant Facial Expression Recognition," *Image Processing, IEEE Transactions on*, vol. 24, pp. 189-204, 2015.

[19] S. Eleftheriadis, O. Rudovic, and M. Pantic, "View-constrained latent variable model for multi-view facial expression classification," in *Advances in Visual Computing*, ed: Springer, pp. 292-303, 2014.

[20] C. K. Williams and C. E. Rasmussen, "Gaussian processes for machine learning," *the MIT Press,* 2006.

[21] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1. 0: A MATLAB Toolbox for Speaker Recognition Research," *Speech and Language Processing Technical Committee Newsletter,* 2013.