Language Recognition With Locality Preserving Projection

Jinchao Yang, Xiang Zhang, Li Lu, Jianping Zhang, Yonghong Yan

ThinkIT Speech Lab, Institute of Acoustics
Chinese Academy of Sciences
Beijing, P.R.China
{yangjinchao, xzhang, llu, jzhang, yonghong.yan}@hccl.ioa.ac.cn

***Abstract* - In this paper, we introduce locality preserving projection (LPP) to language recognition under the support vector machine (SVM) framework. The success of the use of total variability in language recognition shows that the global structure and linear manifold preserve discriminative language dependent information. The proposed LPP language recognition system believes the local structure and nonlinear manifold also contain discriminative language dependent information. Experiment results on 2007 National Institute of Standards and Technology (NIST) language Recognition Evaluation (LRE) databases show LPP language recognition system combining total variability language recognition system gains relative improvement in EER of 11.7% and in minDCF of 9.6% comparing to total variability language recognition system in 30-second tasks, and further improvement is obtained combining with state-of-the-art systems. It leads to gains of 13.8% in EER and 20.2% in minDCF compare with the performance of the combination of the MMI and the GMM-SVM systems.**

***Index Terms*—** language recognition, language total variability, PCA, LDA, LPP, SVM,

I. INTRODUCTION

The aim of language recognition is to determine the language spoken in a given segment of speech. Phoneme recognizer followed by language models (PRLM) and parallel PRLM (PPRLM) approaches that use phonotactic information have shown very successful performance [1][2]. In PPRLM, several tokenizers are used to transcribe the input speech into phoneme strings or lattices [3][4], which are scored by n-gram language models. It is generally believed that phonotactic feature and spectral feature provide complementary cues to each other [1]. The spectral features of speech are collected as independent vectors. The collection of vectors can be extracted as shifted-delta-cepstral acoustic features, and then modeled by Gaussian Mixture Model (GMM). The result was reported in [5]. The approach was further improved by using discriminative train that named Maximum Mutual Information (MMI).

Several studies using SVM in language recognition to form GMM-SVM system [6][7]. SVM as a classifier maps the input feature vector into high dimensional space then separate classes with maximum margin hyperplane. It is important to choose an appropriate SVM feature expansion.

Recently total variability approach has been proposed in speaker recognition [8][9], which uses the factor analysis to define a new low-dimensional space that named total variability space. In this new space, the speaker and the channel variability are contained simultaneously. In ours previous work, we introduce the idea of total variability to language recognition and propose total variability language recognition system. The success of the use of total variability in language recognition show that most of the discriminative language dependent information is captured by low-dimensional subspace.

Actually, total variability method is a classical application of the probabilistic principal component analysis (PPCA) [10]. In our previous work about language recognition with language total variability, we can say that PCA+LDA is used to reduce the dimension of GMM supervector before SVM model. Locality preserving projection (LPP) [11][12] that gains an embedding that preserves local and linear information is different from PCA and LDA which effectively preserve global structure and linear manifold.

In this paper, LPP algorithm is carried out after PCA to a conversation to get the supervector that contain discriminative language dependent information by the local structure and nonlinear manifold. We can call laplacian supervector extraction method.

SVM classifiers are employed to model the laplacian supervector and LDA and diagonal covariance gaussians are used as backend in Language Score Calibration.

This paper is organized as follows: In Section 2, we give a simple review of Support Vector Machines and total variability language recognition system. Section 3 shows the laplacian algorithmic procedure. In Section 4, the proposed language recognition system is presented in detail. corpora and evaluation are given in Section 5. Section 6 gives out experi-

 ISBN: 978-1-61208-127-4

mental result. Finally, we conclude in Section 7.

II. BACKGROUND

A. Support Vector Machines

An SVM [13] is a two-class classifier constructed from sums of a kernel function $K(\cdot,\cdot)$:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(\mathbf{x}, \mathbf{x_i}) + d \quad (1)$$

where N is the number of support vectors, t_i is the ideal output, α_i is the weight for the support vector x_i, $\alpha_i > 0$ and $\sum_{i=1}^{N} \alpha_i t_i = 0$. The ideal outputs are either 1 or -1, depending upon whether the corresponding support vector belongs to class 0 or class 1. For classification, a class decision is based upon whether the value, $f(x)$, is above or below a threshold.

The kernel $K(\cdot,\cdot)$ is constrained to have certain properties (the Mercer condition), so that $K(\cdot,\cdot)$ can be expressed as

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^{'} \phi(\mathbf{y}) \quad (2)$$

where $\phi(x)$ is a mapping from the input space (where $\mathbf{x}$ lives) to a possibly infinite dimensional SVM expansion space. We refer to the $\phi(x)$ as the SVM features.

B. Language Total Variability

In total variability speaker recognition, the factor analysis is used to define a new low-dimensional space that named total variability space and contains the speaker and the channel variability simultaneously. Then, the intersession compensation can be carried out in low-dimensional space. We define language total variability space.

a. Language Total Variability Space Estimation

There is only one difference between total variability space T estimation and eigenvoice space estimation in speaker recognition [9]. All the recordings of speaker are considered to belong to the same person in eigenvoice estimation, however, in total variability space estimation, a given speaker's entire set of utterances are regarded as having been produce by different speakers. We suppose that different conversation from one language is produced by different languages.

For a given conversation, the language and variability dependent supervector is denoted in equation (4).

$$M = m_{ubm} + Tw \quad (3)$$

where m_{ubm} is the UBM supervector, T is language total variability space, and the member of the vector w are total factor. We could call w language total factor vector. We can think the language total factor vector model a new feature extractor that project a conversation to a low rank space T to get a language and variability dependent language total factor vector w.

b. Intersession Compensation

After the feature extractor, the intersession compensation can be carried out in low-dimensional space. We use the Linear Discriminant Analysis approach (LDA) to intersession compensation. All the language total factor vector of the same language are think as the same class.

$$w^* = Aw \quad (4)$$

By LDA transformation in equation (4), the language total factor vector w is projected to new axes that maximize the variance between language and minimizing the intra-class variance. The matrix A is trained using the dataset show in session 5 and is contained of the more larger eigenvectors of equation (5).

$$S_b \nu = \lambda S_w \nu \quad (5)$$

where λ is the diagonal matrix of eigenvalues. The matrix S_b is the between class covariance matrix and S_w is the within class covariance matrix.

III. LAPLACIAN ALGORITHM PROCEDURE WITH LOCALITY PRESERVING PROJECTION

Since total variability method is a classical application of the probabilistic principal component analysis (PPCA). In our previous work about language recognition with language total variability, we can say that PCA+LDA is used to reduce the dimension of GMM supervector before SVM model. As a type of PCA, the total variability method does not need language information. And PCA seeks directions that are efficient for representation. LDA seeks directions that are efficient for discrimination. PCA+LDA that aims to preserve the global structure and linear manifold is successful for language recognition, then the local structure and nonlinear manifolds may be useful to language recognition.

Though LPP is still a linear technique, it seems to reserve important aspects of the intrinsic nonlinear manifold structure by preserving local structure. The algorithmic procedure in this paper is formally stated below.

A. PCA projection

In session 2.2, a conversation is projected to language total variability space to get a language and variability dependent language total factor vector w. Actually, it is a PCA projection. In this paper, LPP is used after PCA projection.

B. Constructing the nearest-neighbor gragh

Let G denote a gragh with m nodes. We put an edge between nodes i and j while i is among k nearest neighbors of j or j is among k nearest neighbors of i.

 ISBN: 978-1-61208-127-4

C. choosing the weights

If nodes i and j are connected, let

$$S_{ij} = e^{-\frac{(w_i - w_j)^2}{t}} \quad (6)$$

The justification for this choice of weights can be traced back to [14].

D. eigenmap

Compute the eigenvectors and eigenvalues for generalized eigenvector problem:

$$WLW^T a = \lambda WDW^T a \quad (7)$$

where D is a diagonal matrix whose entries are column sums of S, $D_{ij} = \sum_j S_{ji}$. $L = D - S$ is the Laplacian matrix. The ith row of matrix W is w_i. Let $a_0, a_1, ..., a_{l-1}$ be the solution of (7), ordered according to their eigenvalues, $0 \leq \lambda_0 \leq \lambda_0 \leq ... \leq \lambda_{l-1}$. Thus, the embedding is as follows:

$$w_i \longrightarrow y_i = A^T w_i, A = (a_0, a_1, ..., a_{l-1}) \quad (8)$$

where y is a l dimensional vector, and A is a transformation matrix.

IV. THE PROPOSED LANGUAGE RECOGNITION SYSTEM

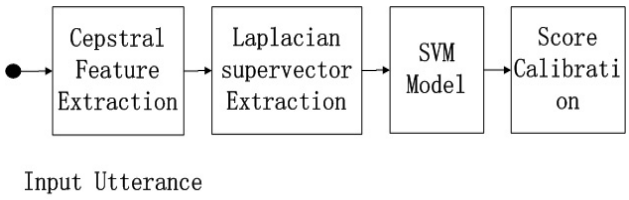

Fig. 1. The Proposed Language recognition System

Figure 1 show the frame of the proposed Language recognition system.

A. MSDC Feature Extraction

The MSDC feature in the system is 7 MFCC coefficient concatenated with SDC 7-1-3-7 feature, which are in total 56 dimension coefficients each frame. MSDC feature refers to this 56 demension feature in my system. Nonspeech frames are eliminated after speech activity detection, then 56 dimension MSDC feature are Extraction. Then feature warping and cepstral variance normalization are applied on the previously extracted MSDC feature which results that each feature is normalized to mean 0 and variance 1.

B. Laplacian supervector Extraction

In our system, we use MSDC feature after compensation of channel factors. Firstly, total variability spaces are estimated as session 2.2. MSDC feature, UBM and Language-independent Total variability space T are need as equation (3) in language total factor vector extraction (actually it is a PCA Projection). Then LPP transformation matrix is learned as session 3. The embedding is as follows to each GMM supervector x:

$$x \longrightarrow y = A^T x \quad (9)$$

$$A = A_{PCA} A_{LPP} \quad (10)$$

where A_{PCA} denote the transformation matrix of PCA as session 2.2. And A_{LPP} denote the transformation matrix of LPP, while the algorithmic procedure is in session 3. We call A Laplacian transformation matrix.

C. SVM Model and Language Score Calibration

Our experiments are implemented using the SVMTorch [15] with a linear inner-product kernel function.

Calibrating confidence scores in mutiple-hypothesis language recognition has been studied in [16]. We should estimate the posterior probability of each hypotheses and make a maximum a posterior decision. In standard SVM-SDC system [7], log-likelihood ratios (LLR) normalization is applied as a simple backend process and is useful. Suppose $S = [S_1 ... S_L]^t$ is the vector of L relative log-likelihoods from the L target languages for a particular message. Considering a flat prior, a new log-likelihood normalized score S'_i is denoted as:

$$S'_i = S_i - log(\frac{1}{L-1} \sum_{j \neq i} e^{S_j}) \quad (11)$$

A more complex full backend process is given [7] [17], LDA and diagonal covariance gaussians are used to calculate the log-likelihoods for each target language and achieve improvement in detection performance.

In this paper, the two backend processes are used in language recognition system. Experiments also show the similar conclusion that the LDA and diagonal covariance gaussians backend process is superior over log-likelihood ratios normalization.

V. CORPORA AND EVALUATION

The experiments are done using the NIST LRE 2007 evaluation database. There are 14 target languages in corpora used in this paper: Arabic, Bengali, Chinese, English, Farsi, German, Hindustani, Japanese, Korean, Russian, Spanish, Tamil, That and Vietnamese. The task of this evaluation was to detect the presence of a hypothesized target language for each test utterance. The Training data was primarily from Callfriend corpora, Callhome corpora, mixer corpora, OHSU corpora, OGI corpora and LRE07Train. The development data

 ISBN: 978-1-61208-127-4

consist of LRE03, LRE05, LRE07Train. We use equal error rate (EER) and the minimum decision cost value (minDCF) as metrics for evaluation.

VI. EXPERIMENTS

Firstly, total variability Language recognition System (PCA+LDA) is experimented, then export to Laplacian Language recognition System (PCA+LPP).

Table 1. Results of the MMI system, GMM-SVM system, the total variability and proposed Laplacian language recognition systems on the NIST LRE07 30s corpus.

System	EER	MinDCF
MMI (a)	3.62	3.78
GMM-SVM (b)	2.65	2.61
total variability (c)	3.15	2.61
Laplacian (d)	3.29	2.83

In Table 1, we give the performance of the MMI, the GMM-SVM, the total variability and proposed Laplacian language recognition systems on NIST 2007 language recognition Evaluation 30s corpus after score backend. EER and minDCF are observed. With the performance comparison, we can see that total variability and proposed Laplacian language recognition systems achieve performance comparable to that obtained with state-of-the-art approaches, which shows my proposed systems are effective and Laplacian language recognition system indeed contains language information.

Table 2. Score Fusion and Super Join Results of the total variability and proposed Laplacian language recognition systems on the NIST LRE07 30s corpus.

System	EER	MinDCF
total variability (c)	3.15	2.61
Laplacian (d)	3.29	2.83
c+d Score Fusion	2.78	2.36
c+d Super Join	2.87	2.51

Table 2 shows the score fusion and super join results of the total variability and proposed Laplacian language recognition systems. The score fusion leads to gains of 11.7% on EER and 9.6% minDCF compare with the performance of the total variability language recognition systems, And super join gains 8.9% on EER and 3.8% minDCF. The result show Laplacian language recognition system that preserves local and nonlinear information includes different language information comparing to total variability language recognition system that preserves global and linear information.

Table 3. Results of the combination of MMI system and GMM-SVM system, and the combination of the MMI system, GMM-SVM system, total variability system, and Laplacian system on the NIST LRE07 30s corpus.

System	EER	MinDCF
Score Fusion (a+b)	2.47	2.42
Score Fusion(a+b+c)	2.18	2.06
Score Fusion(a+b+c+d)	2.13	1.93

Table 3 shows the results of the combination of the MMI system, the GMM-SVM system, the total variability system and the Laplacian system. EER and minDCF are observed. In language recognition evaluation, MMI and GMM-SVM are primary acoustic system. Usually the combination of the MMI system and the GMM-SVM system is the given performance of acoustic system. It leads to gains of 13.8% on EER and 20.2% minDCF compare with the performance of the combination of the MMI and GMM-SVM systems. Lastly, figure 1 gives DET curves for each system and fusion of each system.

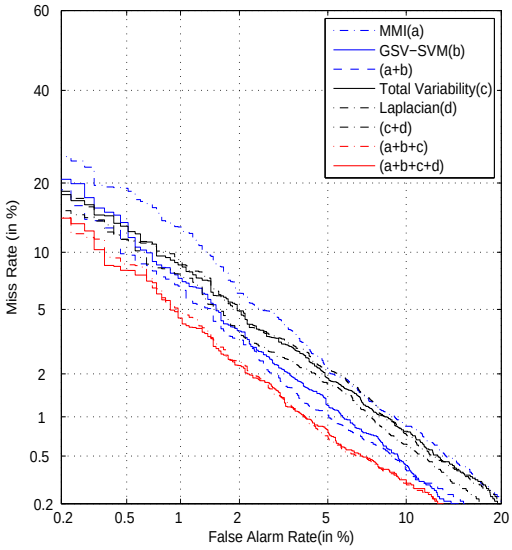

Fig. 2. DET curves for each system and fusion of each system

VII. CONCLUSIONS

In this paper, we propose a new language recognition system by introducing LPP to language recognition. while our previous propose total variability language recognition system show discriminative language dependent information is contained by global structure and linear manifold, the new

 ISBN: 978-1-61208-127-4

language features of Laplacian supervector that preserve local structure and nonlinear manifolds also contain discriminative language dependent information. SVM classifiers are employed to model the new language features and LDA and diagonal covariance gaussians are used as backend in Language Score Calibration. Experiments show that combining two systems LPP and total variability can achieve relative improvement in EER of 11.7% and in minDCF of 9.6% compare to only total variability in 2007 NIST language Recognition Evaluation databases 30-second tasks. Further improvement of relative improvement of 13.8% in EER and 20.2% in minDCF is obtained combining with state-of-the-art systems, comparable with the performance of the combination of the MMI and GMM-SVM systems.

Acknowledgments

This work is partially supported by The National Science and Technology Pillar Program (2008BAI50B00), National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014).

References

[1] M.A. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," in *IEEE International Conference On Acoustics Speech And Signal Processing*. Institute Of Electrical engineers INC (IEE), 1995, vol. 5, pp. 3503–3503.

[2] Y. Yan and E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," in *icassp*. IEEE, 1995, pp. 3511–3514.

[3] J.L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone latices," in *Eighth International Conference on Spoken Language Processing*. ISCA, 2004.

[4] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006*, 2006, pp. 1–6.

[5] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Seventh International Conference on Spoken Language Processing*. Citeseer, 2002.

[6] H. Li, B. Ma, and C.H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, 2007.

[7] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[8] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. International Conferences on Spoken Language Processing (ICSLP)*, 2009, pp. 1559–1562.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *submitted to IEEE Transaction on Audio, Speech and Language Processing*.

[10] M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[11] X. He and P. Niyogi, *Locality preserving projections*, Citeseer, 2005.

[12] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 328–340, 2005.

[13] N. Cristianini and J. Shawe-Taylor, "Support Vector Machines," *Cambridge University Press, Cambridge, UK*, 2000.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 1, pp. 585–592, 2002.

[15] R. Collobert and S. Bengio, "SVMTorch: support vector machines for large-scale regression problems," *The Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.

[16] N. Brummer and D.A. van Leeuwen, "On calibration of language recognition scores," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006*, 2006, pp. 1–8.

[17] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *Eighth European Conference on Speech Communication and Technology*, 2003.

 ISBN: 978-1-61208-127-4