EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Low-dimensional representation of Gaussian mixture model supervector for language recognition

Jinchao Yang*, Xiang Zhang, Hongbin Suo, Li Lu, Jianping Zhang and Yonghong Yan

Abstract

In this article, we propose a new feature which could be used for the framework of SVM-based language recognition, by introducing the idea of total variability used in speaker recognition to language recognition. We consider the new feature as low-dimensional representation of Gaussian mixture model supervector. Thus we propose multiple total variability (MTV) language recognition system based on total variability (TV) language recognition system. Our experiments show that the total factor vector includes the language dependent information; what's more, multiple total factor vector contains more language dependent information. Experimental results on 2007 National Institute of Standards and Technology (NIST) Language Recognition Evaluation (LRE) databases show that MTV outperforms TV in 30 s tasks, and both TV and MTV systems can achieve performance similar to that obtained by state-of-the-art approaches. Best performance of our acoustic language recognition systems can be further improved by combining these two new systems.

Keywords: language recognition, total variability (TV), multiple total variability (MTV), support vector machine, linear discriminant analysis, locality preserving projection

1 Introduction

The aim of language recognition is to determine the language spoken in a given segment of speech. It is generally believed that phonotactic feature and spectral feature provide complementary cues to each other [1,2]. Phone recognizer followed by language models (PRLM) and parallel PRLM (PPRLM) approaches that use phonotactic information have shown very successful performance [2,3]. The acoustic method which uses spectral feature has the advantage that it does not require specialized language knowledge and is computationally simple. This article focuses on the acoustic component of the language recognition systems. 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 [4]. The approach was further improved by using discriminative training that is named maxi-mum mutual information (MMI). Several studies use support vector machine (SVM) in language recognition to form GMM-SVM system [5,6]. In language recognition evaluation, MMI and GMM-SVM are primary acoustic systems.

Recently, total variability approach has been proposed in speaker recognition [7,8], which uses the factor analysis to define a new low-dimensional space that is named total variability space. In contrast to classical joint factor analysis (JFA), the speaker and the channel variability are contained simultaneously in this new space. The intersession compensation can be carried out in low-dimensional space.

Actually, we can consider total variability approach as a classical application of the probabilistic principal component analysis (PPCA) [9]. The factor analysis of the total variability approach can obtain useful information by reducing the dimension of the space of GMM supervectors. That is all utterances could in fact be well represented in a low-dimensional space. We believe useful language information can be obtained by similar front-end processes. Therefore we try to introduce the idea of total variability to language recognition. We estimate the language total variability space by using the

* Correspondence: superyoungking@163.com
Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences, Beijing, P.R. China

dataset shown in Section 5, and we suppose that a given target language's entire set of utterances is regarded as having been belonging to different language. Then, the total factor vector is extracted by projecting an utterance to the language total variability space. As in speaker recognition, intersession compensation can also be performed well on low-dimension total factor vector. In our experiments, two intersession compensation techniques–linear discriminant analysis (LDA) [6] and locality preserving projection (LPP) [10-12]–are used to improve the performance of language recognition.

In some previous studies [13,14], rich information is obtained by using multiple reference models, such as male and female gender-dependent models in speaker recognition. Generally, there are abundant language data for each target language in language recognition, and the number of target languages is limited. Based on TV language recognition system [12,15], we propose MTV language recognition system where we use language-dependent GMMs instead of universal background model (UBM) in the process of language total variability space estimation and total factor vector extraction. Our experiments show that total factor vector (TV system) includes the language dependent information; what's more, multiple total factor vector (MTV system) contains more language dependent information.

This article is organized as follows: In Section 2, we give a simple review of total variability, support vector machines, and compensation of channel factors. In Section 3, we apply total variability in language recognition. In Section 4, the proposed language recognition system is presented in detail. Corpora and evaluation are given in Section 5. Section 6 gives the experimental results. Finally, we conclude in Section 7.

2 Background

2.1 Total variability in speaker recognition

In speaker recognition, unlike in classical joint factor analysis (JFA), the total variability approach defines a new low-dimensional space that is named total variability space, which contains the speaker and the channel variability simultaneously. The total variability approach in speaker recognition relaxes the independent assumption between speaker and channel variability spaces in JFA speaker recognition [16].

For a given utterance, the speaker and channel variability dependent GMM supervector is denoted in Equation (1).

$$M = m_{\text{ubm}} + Tw \tag{1}$$

where m_{ubm} is the UBM supervector, T is total variability space, and the member of the vector w is total factor.

We believe useful language information can be obtained by similar front-end process. Thus we try to apply total variability in language recognition.

2.2 Support vector machines

SVM [17] is used as a classifier after our proposed front-end process in language recognition system. An SVM is a two-class classifier constructed from sums of a kernel function $K(,)$:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(\mathbf{x},\ \mathbf{x_i}) + d \tag{2}$$

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.

2.3 Compensation of channel factors

Compensating the variability from changes in speaker, channel, gender, and environment are the key for the performance of automatic language recognition systems. In our proposed front-end process, the process of an intersession compensation technique in spectral feature domain is still adopted, which has been proposed for speaker and language recognition in [18,19]. The adaptation of the feature vector $\hat{o}^{(i)}(t)$ is obtained by subtracting from the original observation feature a value that is a weighted sum of the intersession compensation offset values.

$$\hat{o}^{(i)}(t) = o^{(i)}(t) - \sum_{m} \gamma_m(t) * U_m * y^{(i)} \tag{3}$$

where $\gamma_m(t)$ is the Gaussian posterior probability of each Gaussian mixture m of the universal background model (UBM) for a given frame of an utterance. U_m and $y^{(i)}$ are about the intersession compensation related to the mth Gaussian of UBM. U_m is intersession subspace and $y^{(i)}$ is channel factor vector. In our proposed language recognition system, we use spectral feature after compensation of channel factors.

3 Applying total variability in language recognition

There is only one difference between total variability space T estimation and eigenvoice space estimation in speaker recognition [8,20]. All the recordings of a speaker are considered as to belong to the same person in the eigenvoice estimation. However, in the total

variability space estimation, a given speaker's entire set of utterances is regarded as having been produced by different speakers. If we suppose that a given target language's entire set of utterances is regarded as having been produced by different languages, a common pool of hidden variables acts as basis factors and represents the utterances from different languages. Then, the process of language total variability space estimation is exactly the same as the process of total variability space estimation and eigenvoice space estimation in speaker recognition. The process is an iterative algorithm [21]. The use of the data which is the only difference is critical. Therefore, we suggest that all utterances of each target language had better be used to estimate language total variability space.

3.1 Language total variability space estimation

For a given utterance, the language and channel variability dependent GMM supervector can also be denoted as Equation (1), because the process of language total variability space estimation is exactly the same as the process of total variability space estimation and eigenvoice space estimation in speaker recognition. We can consider the total factor vector model as a new feature extractor that projects an utterance to a low rank space T to get a language and channel variability dependent total factor vector w. Space estimation can be implemented by an iterative algorithm [21].

3.2 Language-dependent total variability space estimation

In language total variability space estimation, total variability space is estimated relative to UBM, which is language, speaker, channel, gender, and environment independent. Some previous studies [13,14] show that rich information can be obtained by using multiple reference models. These studies suggest the possibility of using language-dependent GMM instead of language-independent UBM in language total variability space estimation. We call language total variability space language-dependent total variability space when the total variability space is related to language-dependent GMM.

First, we train GMM model for each target language. For L target languages, we train a GMM language model for each target language using maximum likelihood (ML) [22]. Then L language-dependent total variability spaces are estimated by using those language dependent GMMs instead of language-independent UBM. An utterance is projected to L different T to get L total factor vectors; as an example, the total factor vector according to Mandarin GMM is illustrated by Equation (4). We combine L total factor vectors to obtain one big multiple total factor vector as Equation (5).

$$M_{\text{mandarin}} = m_{\text{mandarin}} + T_{\text{mandarin}} w_{\text{mandarin}} \tag{4}$$

$$w_{\text{MTV}} = [w_1,\ w_2,\ \ldots,\ w_{\text{mandarin}},\ \ldots, w_L] \tag{5}$$

3.3 Intersession compensation

After the new feature extractor, the intersession compensation can be carried out in low-dimensional space. In our experiment, we use the linear discriminant analysis (LDA) approach and locality preserving projection (LPP) approach for intersession compensation.

3.3.1 Linear discriminant analysis

All of the total factor vectors of the same language are recorded as the same class in linear discriminant analysis.

$$w^* = Aw \tag{6}$$

By LDA transformation in Equation (6), the total factor vector w is projected to new axes that maximize the variance between languages and minimize the intra-class variance. The matrix A is trained by using the dataset shown in Section 5, and the matrix A is contained of the eigenvectors of Equation (7).

$$S_b v = \lambda S_w v \tag{7}$$

where λ is the diagonal matrix of eigenvalues. v is the eigenvector corresponding to the non-zero eigenvalue. The matrix S_b is the between class covariance matrix and S_w is the within class covariance matrix.

3.3.2 Locality preserving projection

Locality preserving projection (LPP) [10,11] is different from LDA which effectively preserves global structure and linear manifold. LPP considers the manifold structure which is modeled by a nearest-neighbor graph. LPP can gain an embedding that preserves local information. In this way, the variability resulting from changes in speaker, channel, gender, and environment may be eliminated or reduced. Thus LPP can be used for intersession compensation.

$$w' = A_{\text{LPP}} w \tag{8}$$

By LPP transformation matrix A_{LPP} in Equation (8), the total factor vector w is projected to w' to preserve local structure of the total factor vector.

First, for training LPP transformation matrix, we construct the nearest-neighbor graph. Let G denotes a graph with m nodes. The ith node corresponds to the total factor vector w_i. 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. In this article, k is set to be 5. If nodes i and j are connected, let

$$E_{ij} = e^{-\frac{(w_i - w_j)^2}{t}} \tag{9}$$

The justification for this choice of weights can be traced back to [23].

Then, we compute the eigenvectors and eigenvalues for generalized eigenvector problem:

$$\mathrm{WLW}^T a = \theta \mathrm{WDW}^T a \tag{10}$$

where D is a diagonal matrix whose entries are column sums of E, $D_{ij} = \sum_j E_{ji}$. $L = D - E$ is the Laplacian matrix. The ith row of matrix W is w_i. Let $a_0, a_1, \ldots, a_{\tau-1}$ be the solution to (10), ordered according to their eigenvalues, $0 \le \theta_0 \le \theta_1 \le \ldots \le \theta_{\tau-1}$. Thus, the LPP transformation matrix is as follows:

$$A_{\mathrm{LPP}} = (a_0, a_1, \ldots, a_{\tau-1}) \tag{11}$$

4 The proposed language recognition system

The proposed TV and MTV language recognition systems contain three main processes, spectral feature extraction, total factor vector extraction, SVM model and language score calibration.

Figure 1 shows the proposed TV and MTV language recognition systems, which contain the three main processes. In Figure 1, the alphabet W is the member of the total factor vector w. N is the dimension of each total factor vector w. GMM1, GMM2, ... , GMML are Gaussian mixture models for each target language.

4.1 Spectral feature extraction

The spectral feature in the system is 7 Mel-frequency cepstral coefficients (MFCC) concatenated with shifted-delta-cpectral (SDC) N-d-p-k feature, where $N = 7$, $d = 1$, $p = 3$, and $k = 7$, which is in total 56-dimension coefficients each frame. This representation is selected based upon prior excellent results with this choice, and the improvement of adding direct coefficients with the C0 coefficient in this feature vector was studied in [24]. In this article, spectral feature refers to this 56-dimension feature. Nonspeech frames are eliminated after speech activity detection and 56-dimension spectral feature is extracted. Then feature warping [25] and cepstral variance normalization are applied to the previously extracted spectral feature such that each feature is normalized to mean 0 and variance 1.

4.2 Total factor vector extraction

In our system, spectral feature after compensation of channel factors is used. First, language total variability space and language-dependent total variability spaces are estimated. Then, we extract total factor vector as shown in Figure 1. In our experiments, the number of mixtures of UBM (or GMM) is 1024, and total variability space T is a rectangular matrix of low rank with dimension 1024*56 by 400. The dimension of w is 400.

The total factor vector w is a hidden variable, and can be obtained as follows [8]:

$$w = (I + T^t \Sigma^{-1} N(u) T)^{-1} T^t \Sigma^{-1} F(\hat{u}) \tag{12}$$

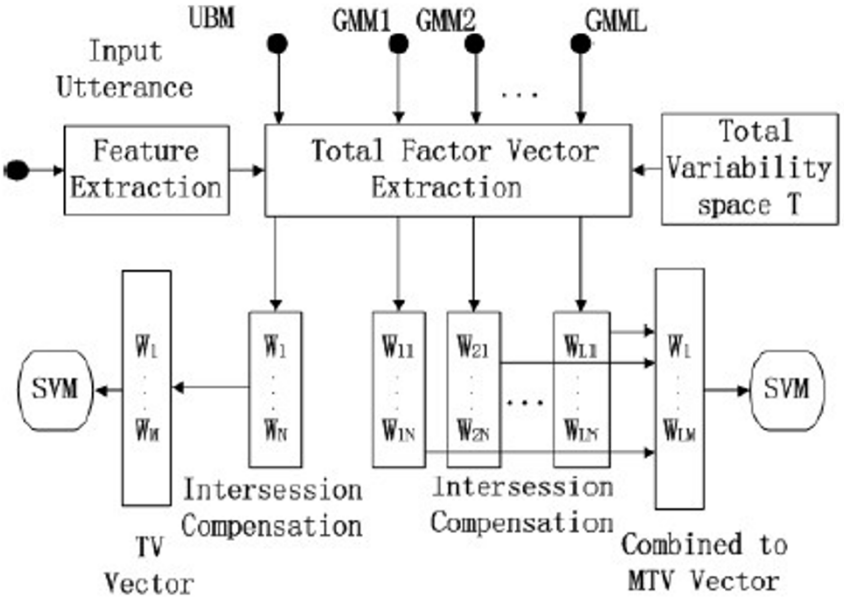

Figure 1 The proposed TV and MTV language recognition systems.

We define $N(u)$ as a diagonal matrix whose diagonal blocks are $N_c I$. $\hat{F(u)}$ is a supervector obtained by concatenating all first-order Baum-welch statistics $\hat{F}_c$ for an utterance u. Σ is a diagonal covariance matrix estimated during factor analysis training [20] and T is language total variability space. N_c and $\hat{F}_c$ are defined as follows:

$$N_c = \sum_{t=1}^{L} P(c|\gamma_t,\ \Omega) \tag{13}$$

$$\hat{F}_c = \sum_{t=1}^{L} P(c|\gamma_t,\ \Omega)(\gamma - m_c) \tag{14}$$

where L is the frames, c is the Gaussian index of C mixture Gaussian components, P (c/y_t, Ω) corresponds to posterior probability of mixture component c generating the vector y_t, and, m_c is the mean of UBM mixture component c.

Multiple total factor vector is extracted with similar method by using language-dependent GMM instead of language-independent UBM and using language-dependent total variability space instead of language total variability space as in Equation (4). Then, the multiple total factor vector w_{MTV} is a combination of w_1, w_2, . . . , w_{mandarin}, . . . , w_L as shown in Figure 1 and Equation (5). Actually, in multiple total variability language recognition system, the combination of total factor vectors is implemented after intersession compensation which is shown in Section 3.3.

4.3 SVM model and language score calibration

Total factor vectors and multiple total factor vectors are used as SVM features in our proposed TV and MTV systems. Our experiments are implemented by using the SVMTorch [26] with a linear inner-product kernel function.

Calibrating confidence scores in multiple-hypothesis language recognition has been studied in [27]. We should estimate the posterior probability of each hypotheses and make a maximum a posterior decision. In standard SVM-SDC system [6], log-likelihood ratios (LLR) normalization is applied as a simple backend process and is useful. Suppose $S = [S_1 \ldots S_L]^t$ is the vector of L relative log-likelihoods from the L target languages for a particular utterance. Considering a flat prior, a new log-likelihood normalized score S_i' is denoted as:

$$S_i' = S_i - \log\left(\frac{1}{L-1}\sum_{j\neq i} e^{S_j}\right) \tag{15}$$

A more complex full backend process is given [6,28], LDA and diagonal covariance Gaussians are used to calculate the log-likelihoods for each target language and achieve improvement in detection performance. This process transforms language scores with LDA, models the transformed scores with diagonal covariance Gaussians (one for each language), and then applies the transform in Equation (15).

In this article, the backend process of the LDA and diagonal covariance Gaussians is used in language recognition system, because the backend process of the LDA and diagonal covariance Gaussians is superior to log-likelihood ratios normalization in our experiments.

5 Corpora and evaluation

The experiments are performed using the NIST LRE 2007 evaluation database. There are 14 target languages in the corpora used in this article: Arabic, Bengali, Chinese, English, Farsi, German, Hindustani, Japanese, Korean, Russian, Spanish, Tamil, Thai, and Vietnamese. The task of this evaluation was to detect the presence of a hypothesized target language for each test utterance. The training data were primarily from Callfriend corpora, Callhome corpora, Mixer corpora, OHSU corpora, OGI corpora, and LRE07Train. The development data consists of LRE03, LRE05, and LRE07Train. We use equal error rate (EER) and the minimum decision cost value (minDCF) as metrics for evaluation.

6 Experiments

First, total variability language recognition system (TV) is experimented, then exports to multiple total variability language recognition system (MTV).

Table 1 shows the results of the MMI system, the GMM-SVM system and the TV and MTV systems with the intersession compensation techniques of LDA and LPP. EER and minDCF are observed. With the performance comparison, it is observed that the two

Table 1 Results of the MMI system, GMM-SVM system and the TV and MTV systems with the intersession compensation techniques of LDA and LPP on the NIST LRE07 30 s corpus

System	EER (%)	MinDCF
MMI (a)	3.62	3.78
GMM-SVM (b)	2.65	2.61
TV(LDA) (c)	3.15	2.61
TV(LPP) (d)	3.29	2.83
TV(LDA+LPP) (e)	2.78	2.36
MTV(LDA) (f)	2.42	2.24
MTV(LPP) (g)	2.83	2.53
MTV(LDA+LPP) (h)	2.32	2.11

intersession compensation techniques of LDA and LPP is effective for TV and MTV systems. The performance is improved obviously when we use LDA and LPP simultaneously. That is models with LDA and models LPP are simultaneously used to score all test utterance. Therefore we regard TV and MTV systems with LDA and LPP simultaneously as our lastly proposed TV and MTV systems. It is observed that the proposed TV and MTV systems achieve performance similar to that obtained by state-of-the-art approaches, which demonstrates that our proposed systems are feasible. Then, we compare the results of TV system to MTV system with the same intersession compensation technique. We can see that the system based on MTV produces better performance than TV. It says multiple total factor vector contain more language-dependent information. In our language recognition systems for NIST 2007 LRE in 30s tasks, the MTV system performs best.

Table 2 shows the results of the combination of the MMI system, the GMM-SVM system, the TV system, and the MTV system, in terms of EER and minDCF. As we know, system fusion can exploit partial error decorrelations among the individual systems allowing for performance gains over the separate systems. In language recognition evaluation, MMI and GMM-SVM are primary acoustic systems. Generally, the combination of the MMI system and the GMM-SVM system is the given performance of acoustic system. Table 1 shows that our proposed TV and MTV systems have been effective. We believe that the TV and MTV systems contain different language information comparing to state-of-the-art systems, because total factor vector and multiple total factor vector are new features for language recognition. Thus we expect the TV and MTV system can benefit the performance of combined system. It leads to a relative improvement of 8.1% in EER and 16.5% in minDCF combining TV system with the MMI and GMM-SVM systems. Further more, we obtain relative improvement of 12.3% in EER and 11.4% in minDCF by adding MTV system to the combined system of the MMI, GMM-SVM, and TV systems. In all, the two systems lead to relative improvement of 19.4% in EER and 26.0% in minDCF comparing to the performance of the combination of the MMI and GMM-SVM systems.

Table 2 Results of the combination of MMI system and GMM-SVM system, and the combination of the MMI system, GMM-SVM system, TV system, and MTV system on the NIST LRE07 30 s corpus

System	EER (%)	MinDCF
Fusion(a+b)	2.47	2.42
Fusion(a+b+e)	2.27	2.02
Fusion(a+b+e+h)	1.99	1.79

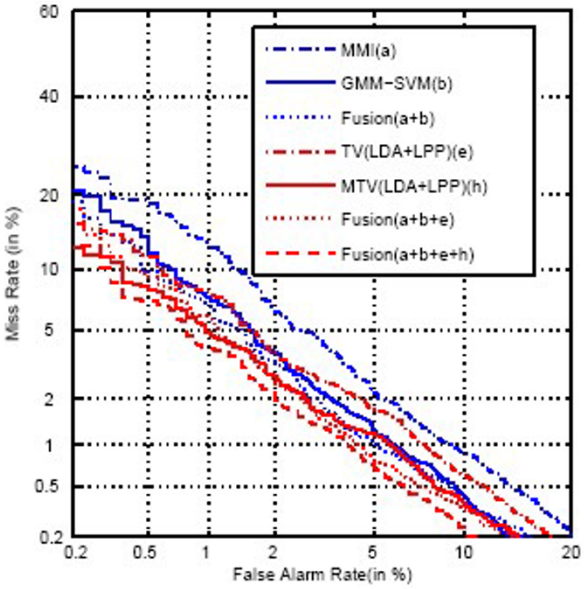

Figure 2 DET curves for each system and fusion of each system.

Figure 2 shows DET curves of the MMI system, GMM-SVM system, the TV system and the MTV system. DET curves of the combination of each system are also shown in Figure 2. It is observed that the relative improvement of language recognition performance is observable with our proposed approaches.

7 Conclusions

In this article, multiple total factor vector are proposed for language recognition based on using total factor vector in language recognition. Our experiments show that total factor vector includes the language dependent information. Further more, multiple total factor vector contains more language dependent information. Comparing to popular acoustic system (MMI and GMM-SVM system) in language recognition, those two new language features contain different language dependent information. We believe it is attractive that our proposed features can improve our best acoustic performance of the combination of the MMI and GMM-SVM systems. In our future study, different approaches of intersession compensation will be carried on the new features.

Acknowledgements

This study was partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275).

Competing interests
The authors declare that they have no competing interests.

Received: 14 May 2011 Accepted: 27 February 2012
Published: 27 February 2012

References

1. PA Torres-Carrasquillo, E Singer, WM Campbell, T Gleason, A McCree, DA Reynolds, F Richardson, W Shen, DE Sturim, "The mitll nist lre 2007 language recognition system", Ninth Annual Conference of the International Speech Communication Association, **1**, Brisbane, Australia, 719–722 (2008)
2. MA Zissman, "Language identification using phoneme recognition and phonotactic language modeling", in *IEEE International Conference On Acoustics Speech And Signal Processing*, vol. 5. Institute Of Electrical engineers INC (IEE). Detroit USA, **5**, 3503-3503, (1995)
3. Y Yan, E Barnard, "An approach to automatic language identification based on language-dependent phone recognition", in *icassp* Detroit USA, IEEE, **5**, 3511–3514 (1995)
4. PA Torres-Carrasquillo, E Singer, MA Kohler, RJ Greene, DA Reynolds, JR Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features", in *Seventh International Conference on Spoken Language Processing*, Citeseer, **1**, 89–92 (2002)
5. H Li, B Ma, CH Lee, "A vector space modeling approach to spoken language identification". IEEE Transactions on Audio, Speech, and Language Processing. **15**(1), 271–284 (2007)
6. WM Campbell, JP Campbell, DA Reynolds, E Singer, PA Torres-Carrasquillo, "Support vector machines for speaker and language recognition". Computer Speech & Language. **20**(2-3), 210–229 (2006). doi:10.1016/j.csl.2005.06.003
7. N Dehak, R Dehak, P Kenny, N Brümmer, P Ouellet, P Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification", in *Tenth Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, **1**, 1559-1562 (2009)
8. N Dehak, P Kenny, R Dehak, P Dumouchel, P Ouellet, "Front-end factor analysis for speaker verification". Audio, Speech, and Language Processing, IEEE Transactions on. **19**(4) (2011)
9. ME Tipping, CM Bishop, "Probabilistic principal component analysis". Journal of the Royal Statistical Society: Series B (Statistical Methodology). **61**(3), 611–622 (1999). doi:10.1111/1467-9868.00196
10. X He, P Niyogi, "Locality preserving projections", in *Advances in neural information processing systems 16: proceedings of the 2003 conference*, Citeseer, **6**, 153-160 (2003)
11. X He, S Yan, Y Hu, P Niyogi, HJ Zhang, "Face recognition using laplacianfaces". IEEE Transactions on Pattern Analysis and Machine Intelligence. **27**, 328–340 (2005)
12. J Yang, X Zhang, L Lu, J Zhang, Y Yan, "Language Recognition With Locality Preserving Projection", The Sixth International Conference on Digital Telecommunications (ICDT 2011), Budapest, Hungary, 46–50 (2011)
13. A Stolcke, SS Kajarekar, L Ferrer, E Shrinberg, "Speaker recognition with session variability normalization based on mllr adaptation transforms". *Audio, Speech, and Language Processing, IEEE Transactions on* **15**(7), 1987–1998 (2007)
14. M Ferras, CC Leung, C Barras, JL Gauvain, "Comparison of speaker adaptation methods as feature extraction for svm-based speaker recognition". *Audio, Speech, and Language Processing, IEEE Transactions on*. **18**(6), 1366–1378 (2010)
15. N Dehak, PA Torres-Carrasquillo, D Reynolds, R Dehak, "Language recognition via ivectors and dimensionality reduction", in *12th Annual Conference of the International Speech Communication Association*. **1**, 857–860 (2011)
16. P Kenny, P Ouellet, N Dehak, V Gupta, P Dumouchel, "A study of interspeaker variability in speaker verification". Audio, Speech, and Language Processing, IEEE Transactions on. **16**(5), 980–988 (2008)
17. N Cristianini, J Shawe-Taylor, *"Support Vector Machines"*, (Cambridge University Press, Cambridge, UK, 2000)
18. F Castaldo, D Colibro, E Dalmasso, P Laface, C Vair, "Compensation of nuisance factors for speaker and language recognition". Audio, Speech, and Language Processing, IEEE Transactions on. **15**(7), 1969–1978 (2007)
19. F Castaldo, S Cumani, P Laface, D Colibro, "Language recognition using language factors", in *Tenth Annual Conference of the International Speech Communication Association*, Brighton, U.K, 176–179 (2009)
20. P Kenny, G Boulianne, P Dumouchel, "Eigenvoice modeling with sparse training data". Speech and Audio Processing, IEEE Transactions on. **13**(3), 345–354 (2005)
21. R Kuhn, JC Junqua, P Nguyen, N Niedzielski, "Rapid speaker adaptation in eigenvoice space". Speech and Audio Processing, IEEE Transactions on. **8**(6), 695–707 (2000). doi:10.1109/89.876308
22. AP Dempster, NM Laird, DB Rubin, "Maximum likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society. Series B (Methodological). **39**(1), 1–38 (1977)
23. M Belkin, P Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering". Advances in neural information processing systems. **1**, 585–592 (2002)
24. L Burget, P Matějka, J Černocký, "Discriminative Training Techniques for Acoustic Language", in *Proceedings of ICASSP*, Toulouse, France, **1**, 209–212 (2006)
25. F Allen, E Ambikairajah, J Epps, "Warped magnitude and phase-based features for language identification", in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, (IEEE Toulouse, France), **1**, 209-204 (2006)
26. R Collobert, S Bengio, "SVMTorch: support vector machines for large-scale regression problems". The Journal of Machine Learning Research. **1**, 143–160 (2001)
27. N Brummer, DA van Leeuwen, "On calibration of language recognition scores", in *IEEE Odyssey 2006:The Speaker and Language Recognition Workshop, 2006*, San Juan, Puerto Rico, **1**, 1-8 (2006)
28. E Singer, PA Torres-Carrasquillo, TP Gleason, WM Campbell, DA Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification", in *Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, **1**, 1345-1348 (2003)

doi:10.1186/1687-6180-2012-47
Cite this article as: Yang *et al.*: **Low-dimensional representation of Gaussian mixture model supervector for language recognition.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:47.