International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.7 (2015), pp.247-254
http://dx.doi.org/10.14257/ijmue.2015.10.7.26

Speaker Recognition via Block Sparse Bayesian Learning

Wei Wang, Jiqing Han*, Tieran Zheng, Guibin Zheng and Mingguang Shao

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
wangwei_hitwh@126.com, jqhan@hit.edu.cn, zhengtieran@hit.edu.cn, zhengguibin@hit.edu.cn, 18369185665@163.com

Abstract

In order to demonstrate the effectiveness of sparse representation techniques for speaker recognition, a dictionary of feature vectors belonging to all speakers is constructed by total variability i-vectors. Each feature vector from unknown utterance is expressed as linear weighted sum of a dictionary. The weights are calculated using Block Sparse Bayesian Learning (BSBL) where the sparsest solution can be obtained. By exploiting the speech signal's block structure and intra-block correlation, the system performance is improved. The experimental results validate that our method outperforms the baseline systems and the system using Orthogonal Matching Pursuit (OMP) algorithm on the typical corpus and realizes the identity validation function.

Keywords: *speaker recognition, sparse representation, Intra-Block Correlation*

1. Introduction

Speaker recognition is one of the most important research fields in the speech technology industry which aims to recognize a person's identity from the characteristics of his voice. This technology is widely used in banking, defense, forensics, video games, and also as front-end of other speech-related tasks like speech recognition. In last few years, sparse signal representation is widespread applied in digital signal processing [1-3]. Originally, the sparse representation was for an efficient presentation and compression of signals at a greatly reduced rate than the standard Shannon-Nyquist rate [4-5]. In recent years, sparse representations based on classifiers have been used in many applications, and experimental results show they achieve the better performance. Sparse representations mainly include two aspects which are the composition of over-complete dictionary and computing of the sparse representation coefficients. In [6-7], the use of the GMM mean supervectors were proposed to develop an over-complete dictionary using all the training speakers for speaker identification and speaker verification. However, sparse representation of large dimension supervector not only requires a large training data where the over-complete dictionary needs that the supervector dimension must be smaller than the number of samples [8], but also requires a large amount of memory space where the over-complete dictionary limit the training sample numbers and slow down the recognition process. Since i-vector has low dimensionality and excellent discriminative capability, Li *et al* proposed the use of i-vectors to develop an over-complete dictionary in [9]. Therefore, in this work, we propose to use i-vectors to construct the over-complete dictionary. In the aspect of computing the sparse representation coefficients, l_1 -minimization is popular, but its global minimum is generally not the sparsest solution. In [10], Tipping *et al* proposed sparse Bayesian learning (SBL) where the global minima of SBL are always the sparsest one. And in SBL [11-12], robust learning rules for automatically estimating values of its regularize are provided such that SBL can achieve good performance. Since many algorithms [13-16] have been proposed that signals often contain some kind of structures, exploiting special block structures in the sparse signal

ISSN: 1975-0080 IJMUE

where only a few blocks are nonzero and the intra-block correlation is helpful in improving the performance [11]. Therefore, in this work, we propose to use Block Sparse Bayesian Learning (BSBL) to estimate the sparse representation coefficients by exploiting the block structure and the intra-block correlation.

2. Sparse Representation Classification

2.1. I-Vector Feature Extraction

Currently, one of the challenges in speaker recognition is channel variability between the training and testing data [17-18]. In [19], Dehak *et al* proposed the i-vectors. The i-vectors extraction could be as a compression process which reduces the dimensionality of speech and channel supervectors. Given an utterance, the speaker and channel dependent GMM mean supervector defined by (1) is written as follows:

$$M = m + T\omega \tag{1}$$

Where m is the UBM supervector, T is called Total Variability matrix that is a rectangular matrix of low rank and ω is i-vector. In i-vectors, the total variability space contains the speaker and channel effects simultaneously, the speaker and channel dependent GMM supervector M is projected in the low rank space T, and the low dimensionality vector ω is got. Channel compensation is applied based on within-class covariance normalization (WCCN) [20] and linear discriminant analysis (LDA) [19].

2.2. Dictionary Composition

Supposed that there are K speakers, and each speaker has a set of N i-vectors extracted. Each i-vector contains a m-dimensional vector. Let

$$D_k = [d_{k1}\, d_{k2} \ldots d_{kn}\, d_{kN}] \in R^{m \times N} \tag{2}$$

be a $m \times N$ matrix of i-vectors of the kth speaker, where the column $d_{kn} = [d_{kn_1} d_{kn2} \ldots d_{kn_m}]^T$ denotes the m-dimensional i-vector of the nth number belonging to the kth speaker. A dictionary D for sparse representation can be constructed by concatenating i-vectors of all the K speakers

$$\begin{aligned} D &= [D_1\, D_2 \ldots D_k \ldots D_K] \in R^{m \times KN} \\ &= [d_{11} \ldots d_{1N} | d_{21} \ldots d_{2N} | \ldots | d_{k1} \ldots d_{kN} | \ldots | d_{K1} \ldots d_{KN}] \end{aligned} \tag{3}$$

A test i-vector $y \in R^m$ can be expressed as a linear weighted sum of columns of dictionary D as

$$y = \sum_{k=1}^{K} \sum_{n=1}^{N} d_{kn} \alpha_{kn} \tag{4}$$

where α_{kn} is the weighted associated with the column d_{kn}.

The equation (4) can be defined the matrix form as

$$y = D\alpha \tag{5}$$

where $\alpha = [\alpha_{11} \ldots \alpha_{1N} \,|\, \alpha_{21} \ldots \alpha_{2N} \,|\, \ldots |\, \alpha_{k1} \ldots \alpha_{kN} \,|\, \ldots |\, \alpha_{K1} \ldots \alpha_{KN}]^T$

In order to represent a test i-vector y as a linear combination of columns D, we select the BSBL algorithm to obtain an approximate sparse weight vector α.

2.3. Speaker Recognition via BSBL

The sparse representation coefficient vector α can be viewed as a concatenation of K non-overlapping blocks:

$$\alpha = [\underbrace{\alpha_{11}...\alpha_{1N}}_{\alpha_1^T}\big|\ \alpha_{21}...\alpha_{2N}\ \big|...\big|\alpha_{k1}...\alpha_{kN}\big|...\big|\underbrace{\alpha_{K1}...\alpha_{KN}}_{\alpha_K^T}]^T \quad (6)$$

each block $\alpha_k \in R^{N\times 1}$ is ass α_1^T to be mutually indepe α_K^T and satisfy a parameterized multivariate Gaussian distribution:

$$p(\alpha_k;\gamma_k,B_k) \sim N(0,\gamma_k B_k) \qquad k=1,...,K$$

with the unknown parameters γ_k and B_K. The nonnegative parameter γ_k controls the block-sparsity of α. When $\gamma_k = 0$, the α_k becomes zero, and most γ_k tend to be zero [10]. A positive definite matrix $B_K \in R^{N\times N}$ captures the correlation structure of the α_k. Besides, the noise vector is assumed to be independent and satisfy $p(v;\lambda) \sim N(0,\lambda I)$, where λ is a positive scalar. The prior of α is $p(\alpha;\{\gamma_k,B_k\}_k) \sim N(0,\sum_0)$ with the assumption that blocks are mutually uncorrelated. Using the Bayesian rule, the posterior of α is $p(\alpha|y;\lambda,\{\gamma_k,B_k\}_{k=1}^K) \sim N(\mu_\alpha,\sum_\alpha)$ which is also Gaussian, with the mean $\mu_\alpha = \sum_0 \phi^T(\lambda I + \phi\sum_0\phi^T)^{-1}y$ and the covariance $\sum_\alpha = (\sum_0^{-1} + \frac{1}{\lambda}\phi^T\phi)^{-1}$.

The parameters $\lambda,\{\gamma_k,B_k\}_{k=1}^K$ can be estimated by a Type Ⅱ maximum likelihood procedure [10], which is equivalent to minimizing the cost function

$$L(\lambda,\{\gamma_k,B_k\}_{k=1}^K) \overset{\Delta}{=} \log|\lambda I + \phi\sum_0\phi^T| + y^T(\lambda I + \phi\sum_0\phi^T)^{-1}y \quad (7)$$

So given the parameters $\lambda,\{\gamma_k,B_k\}_{k=1}^K$, the Maximum-A-posterior (MAP) of α can be got by the mean of the posterior

$$\overline{\alpha} = \sum_0\phi^T(\lambda I + \phi\sum_0\phi^T)^{-1}y \quad (8)$$

In evaluation, if the given test i-vector sample y belongs to the *k*th speaker, the residue that are associated with the *k*th speaker should be smallest. The given test i-vector sample y is approximated as $y_k = D\delta_k(\alpha)$. y is assigned to the object class *k* that gave the smallest residual between y and y_k

$$k = \arg\min\| y - y_k \|_2 = \arg\min\| y - D\delta_k(\alpha)\|_2 \quad (9)$$

Where $\delta_k(\alpha)$ is a vector that selected the only nonzero coefficients associated with the *kth* class as shown in

$$\delta_k(\alpha) = [0...0\,|...|\,\alpha_{k1}...\alpha_{kN}\,|...|\,0...0]^T \quad (10)$$

3. Experimental Evaluations

In this study, the i-vector as features is used to represent the speaker characteristic of the train set and the test set. The over-complete dictionary is composed of the normalized i-vectors (with unit l_2 norm) of training utterances. The i-vector of a test utterance is represented as a linear weighted sum of this over-complete dictionary. And the associated weight vector is obtained using BSBL algorithm. If the reconstruction residual that are associated with the speaker is the smallest, the test belongs to the speaker. The flow chart of the overview of the system architecture is shown in Figure 1.

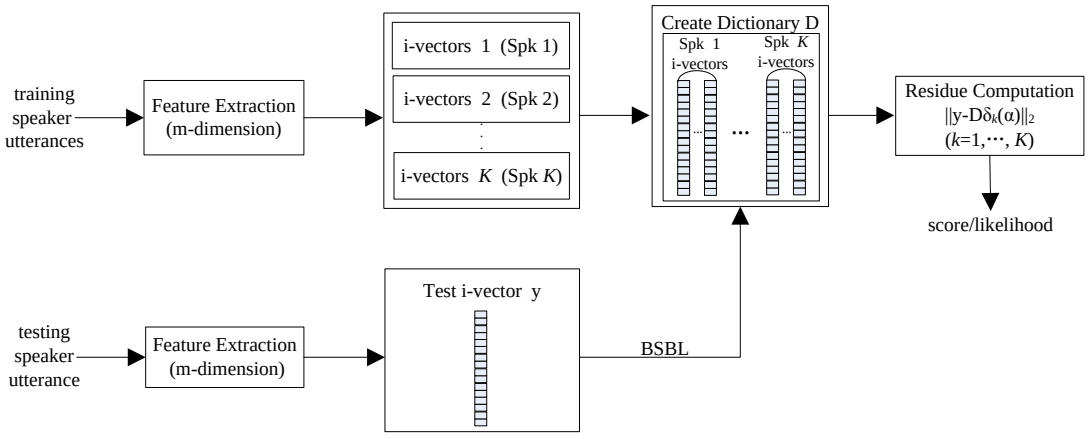

Figure 1. Architecture of the System

In order to evaluate the performance of the proposed method, experiments are carried out on NIST SRE 2003 [21] database. In the female set of NIST SRE 2003, the train set consists of 207 target speaker speeches and each of them is approximately 2 minutes. The test set has 1795 true trials and 17950 false trials. And in the male set of NIST SRE 2003, the train set is consisted of 149 target speaker speeches and each of them is approximately 2 minutes. The test set has 1345 true trials and 13450 false trials. All the speech files which are collected over telephone channels are the wave format at the sample frequency of 8 kHz and quantized with 16 bits. Speech streams are windowed into a sequence of short-term frames (20 ms long) with 10ms overlapped data. Furthermore, the speech files use 34-dimensional MFCC (16+log_energy, appended with their first deltas) with the cepstral mean subtraction (CMS) [22] and the feature warping [22] in order to remove any factors related to the recording conditions.

The baseline system is a gender-dependent Gaussian mixture model and universal background model (GMM-UBM) in which OGI corpus and NIST SRE 2003 training corpus are used to train a UBM with 1024 mixtures. And the current state of the baseline system i-vector is implemented based on cosine distance scoring (i-CDS) [20]. For i-vector, the total variability matrix composed of 127 total factors was trained with OGI and NIST SRE 2003. The dimensionality of i-vector is 127. In NIST SRE 2003 speaker recognition evaluation, each test utterance includes 11 claimants. Every claimant is extracted 10 i-vectors. The 127-dimensional i-vectors from all 11 claimants are concatenated to form a over-complete dictionary which contains 110 atoms. The performance of the system is measured by Equal Error Rate (EER). The results in the female set of NIST SRE 2003 database are shown in Figure 2. The solid and dotted lines are used to describe Detection Error Tradeoffs (DETs) of GMM-UBM (EER with 10.97%) and i-CDS (EER with 6.7%) of the baseline systems respectively, and the dashed line is used to describe the DETs of the BSBL framework (EER with 6.14%). And the results in the male set of NIST SRE 2003 database are shown in Figure 3. The solid and dotted lines are used to describe DETs of GMM-UBM (EER with 8.96%) and i-CDS (EER with 6.34%) of the baseline systems respectively, and the dashed line is used to describe the DETs of the BSBL framework (EER with 6.03%). It can be seen that BSBL framework significantly outperforms the baseline systems (GMM-UBM and i-CDS).

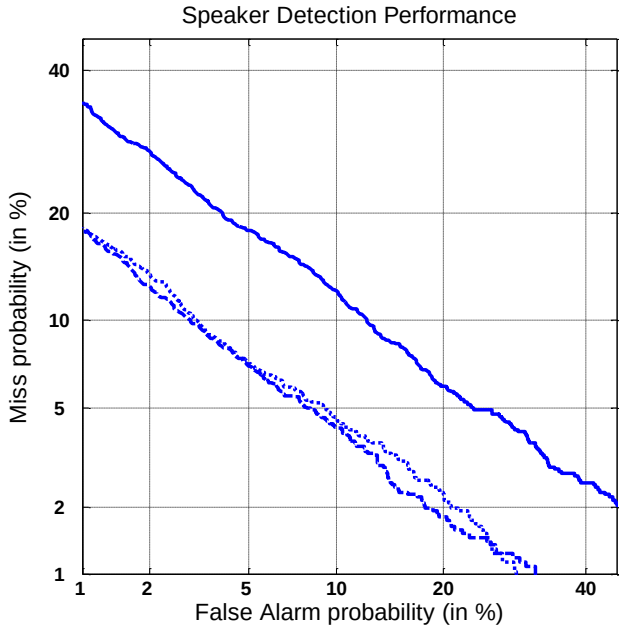

Figure 2. DETs of the Baseline System (The Solid and Dotted Lines) and BSBL Framework (The Dashed Line) in the Female set of NIST SRE 2003 database.

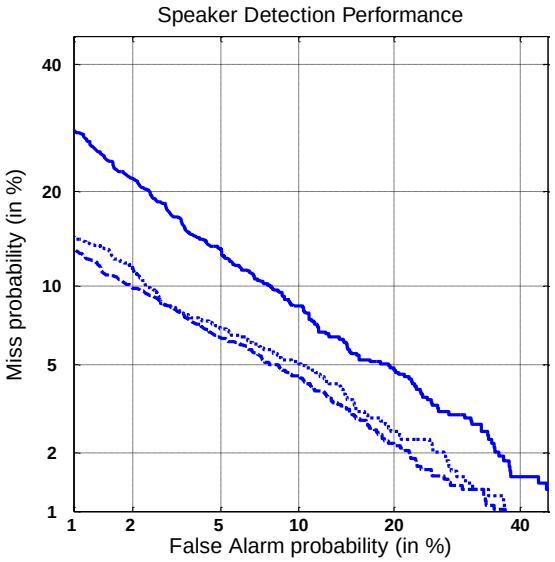

Figure 3. DETs of the Baseline System (The Solid and Dotted Lines) and BSBL Framework (The Dashed Line) in the Male Set of NIST SRE 2003 Database.

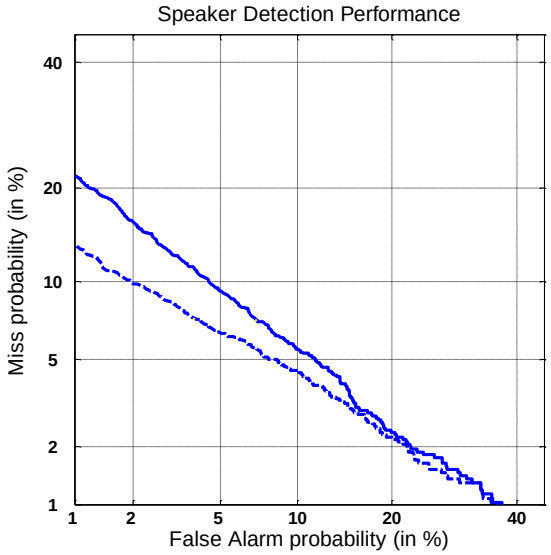

Figure 4. Performance comparision between exploiting the intra-block correlation and ignoring the intra-block correlation in the female set of NIST SRE 2003 database.

To clarify a benefit to exploiting intra-block correlation, the results in the female set of NIST SRE 2003 are shown in Figure 4, where the solid and dotted lines are used to describe DETs of OMP (EER with 7.99%) and BSBL (EER with 6.14%) of the frameworks respectively. And the results in the male set of NIST SRE 2003 are shown in Figure 5, where the solid and dotted lines are used to describe DETs of OMP (EER with 7.1%) and BSBL (EER with 6.03%) of the frameworks respectively. It can be seen that BSBL framework of exploiting the intra-block correlation significantly outperforms the framework of ignoring the intra-block correlation.

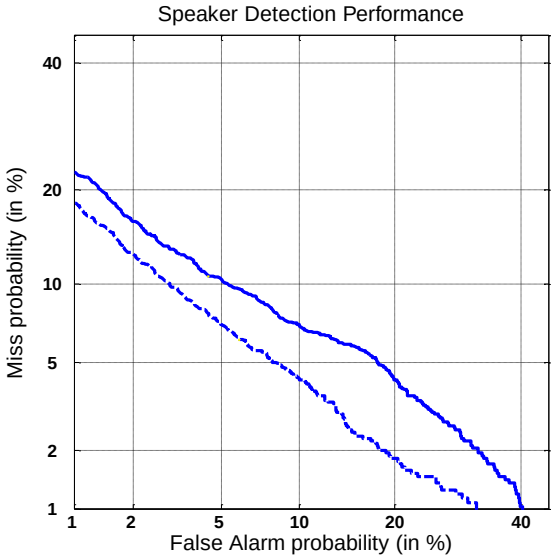

Figure 5. Performance Comparison between Exploiting the Intra-Block Correlation and Ignoring the Intra-Block Correlation in the Male set of NIST SRE 2003 Database.

4. Conclusions

In this paper, we have proposed the sparse representation techniques for text-independent speaker recognition. The i-vector feature of a given testing utterance is expressed as linear combination of a set of i-vector features of training utterances. This involves solving underdetermined system of linear equations for a sparse solution. We have used BSBL algorithm for obtaining the sparse solution. This algorithm exploits intra-block correlation in signals and thereby improves performance. It is observed that the performance of BSBL algorithm outperforms the basic sparse representation algorithm.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 91120303) and the Ph.D. Programs Foundation of Ministry of Education of China (No. 2011230-2110042).

References

[1] E. J. Cande`s, "Editors. Compressive sampling," Proceedings of the 2th International Congress of Mathematicians, Madrid, Spain, August 22-30, **(2006)**.
[2] A. Y. Yang, M. Gastpar, R. Bajcsy and S. S. Sastry, "Proceedings of the IEEE," vol. 98, **(2010)**, pp. 1077.
[3] K. Huang and S. Aviyente, "Advances in Neural Information Processing Systems," vol. 19, **(2007)**, pp. 609.
[4] E. J. Cande`s, J. Romberg and T. Tao, "IEEE Transactions on Information Theory," vol. 52, **(2006)**, pp. 489.
[5] R. G. Baraniuk, "IEEE Signal Processing Magazine, vol. 24, **(2007)**.
[6] I. Naseem, R. Togneri and M. Bennamoun, "Sparse Representation for Speaker Identification," Proceedings of the 20th International Congress of Pattern Recognition (ICPR), Istanbul, Turkey, August 23-26, **(2010)**.
[7] Kua J. M. K., Ambikairajah E., Epps J. and Togneri R., "Speaker Verification Using Sparse Representation Classification," Proceedings of the 36th International Congress of Acoustics, Speech and Signal Processing, Prague, Czech, May 22-27, **(2010)**.
[8] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Pattern Analysis and Machine Intelligence," vol. 31, **(2008)**, pp. 210.
[9] M. Li, X. Zhang, Y. Yan and S. Narayanan, "Speaker Verification Using Sparse Representations on Total Variability I-vectors, Proceedings of the 12th International Congress of the International Speech Communication Association (Inter speech), Florence, Italy, August 28-31, **(2011)**.
[10] M. E. Tipping, "The journal of machine learning research," vol. 1, **(2001)**, pp. 211.
[11] Z. Zhang and B. D. Rao, "IEEE Transactions on Signal Processing," vol. 61, **(2013)**, pp. 2009.
[12] Z. Zhang and B. D. Rao, "IEEE Transactions on Signal Processing," vol. 5, **(2011)**, pp. 912.
[13] Y. C. Eldar and M. Mishali, "IEEE Transactions on Information Theory," vol. 55, **(2009)**, pp. 5302.
[14] R. G. Baraniuk, V. Cevher and M. F. Duarte, "EEE Transactions on Hegde," Information Theory, vol. 56, **(2010)**, pp. 1982.
[15] M. Yuan and Y. Lin, "Journal of the Royal Statistical Society: Series B (Statistical Methodology),"vol. 68, **(2006)**, pp. 49.
[16] P. Zhao, G. Rocha and B. Yu, "The Annals of Statistics," vol. 37, **(2009)**, pp. 3468.
[17] A. Solomonoff, C. Quillen and W. M. Campbell, "Channel Compensation for SVM Speaker Recognition," Proceedings of the International Congress of The Speaker and Language Recognition Workshop(Odyssey), Toledo, Spain, May 31-June 3, **(2004)**.
[18] A. Solomonoff, W. M. Campbell and I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," Proceedings of the 29th International Congress of Acoustics, Speech and Signal Processing, Philadelphia, USA, March 19-23, **(2005)**.
[19] N. Dehak, P. Kenny and R. Dehak, "IEEE Transactions on Audio, Speech, and Language Processing," vol. 19, **(2011)**, pp. 788.
[20] N. Dehak, P. Kenny and R. Dehak, "Editors. Support Vector Machines and Joint Factor Analysis for Speaker Verification," Proceedings of the 34th International Congress of Acoustics, Speech and Signal Processing, Taipei, Taiwan, April 19-24, **(2009)**.
[21] M. Przybocki and A. Martin, "The NIST year 2003 speaker recognition evaluation plan," http://www.nist.gov/tests/spk/2003/index.htm, **(2003)**.
[22] T. Kinnunen and H. Li, "Speech Communication," vol. 52, **(2010)**, pp.12.

Authors

Wei Wang, she received a B.S. degree in School of Computer Science and Technology from Harbin Institute of Technology 2002, and M.S. degree in Software Engineering from Harbin Institute of Technology 2004.

Jiqing Han, he received the B.S., M.S. in electrical engineering, and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1987, 1990, and 1998, respectively. Currently, he is a Professor of the School of Computer Science and Technology, Harbin Institute of Technology. His research fields include speech signal processing and audio information processing. He is a member of IEEE, Vice Chairman of Society of speech processing, Association for Chinese information processing, Vice Chairman of Standing Committee of National Conference on Man-Machine speech Communication, China, member of the editorial board of Journal of Chinese Information Processing, and member of the editorial board of the Journal of Data Acquisition & Processing.