Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Complementary Combination in I-vector Level for Language Recognition

Zhi-Yi Li, Wei-Qiang Zhang, Liang He, Jia Liu

Department of Electronic Engineering
Tsinghua University, Beijing, China 100084
lizhiyi06@mails.tsinghua.edu.cn, {wqzhang, heliang, liuj}@tsinghua.edu.cn

Abstract

Recently, i-vector based technology can provide good performance in language recognition (LRE). From the viewpoint of information theory, i-vectors derived from different acoustic features can contain more useful and complementary language information. In this paper, we propose an effective complementary combination for two kinds of i-vectors. One is derived from the commonly used short-term spectral shifted delta cepstral (SDC) and the other from a novel spectro-temporal time-frequency cepstrum (TFC). In order to overcome the curse of dimension and to remove the redundant information in the combined i-vectors, we use principal component analysis (PCA) and linear discriminant analysis (LDA) and evaluate their performances, respectively. For classification, cosine distance scoring (CDS) and support vector machine (SVM) are applied to the new combined i-vectors. The experiments are performed on the NIST LRE 2009 dataset, and the results show that the proposed method can effectively improve the better performance than baseline by EER reducing 1% for 30 s duration and 2.3% for both 10 s and 3 s.

Index Terms— i-vector combination, SDC, TFC, PCA, LDA, language recognition

1. Introduction

Language recognition (LRE) refers to automatically recognize the language from a speech utterance. It has applications in many areas, such as multi-lingual speech-related services, information security, etc.

Over the past years, the well-performed systems developed in LRE can be simply classified as the phonotactic systems and the acoustic systems. The former ones typically focus on the phones and the frequency of the phone sequences observed in each target language, while the latter ones mainly base on the spectral characteristics of each language. Generally speaking, many of the acoustic LRE systems are founded on the same algorithms as the speaker recognition (SRE) systems, such as Gaussian mixture models (GMM) [1], support vector machines (SVM) [2], joint factor analysis (JFA) [3], etc. On the other hand, many well-performed technologies in SRE can always show the identically excellent performance in LRE [4]. Recently, i-vector based technology can provide the better performance than JFA in SRE [4], and many researches have reported this advantage in LRE [5, 6]. In i-vector based systems, the fixed length low-dimension i-vectors are extracted by estimating the latent variables from each utterance based on the factor analysis algorithm like JFA and then used to be the inputs for the classifier.

In the meanwhile, it has been proved yet that different acoustic features can reflect complementary discriminant information of languages. In addition, it has large influence on the performance of classifier. Even though various high-level or other features have been studied, acoustic features based on spectrum still outperform the others very well and are the most widely used in practice. In LRE, shifted delta cepstral (SDC) and time-frequency cepstrum TFC [8] have been considered as two effective and complementary well-performed acoustic features.

From the viewpoint of information theory, complementary use of i-vectors derived from different acoustic features can also contain more useful discriminant information. In this paper, we explore more about the complementary combination method in i-vector level. At first, multiple complementary i-vectors extracted from different acoustic features are simply concatenated to be a new higher-dimensional vector. Then, in order to avoid the high dimension and redundant information, unsupervised principal component analysis (PCA) and supervised linear discriminant analysis (LDA) are used respectively and their performances are evaluated. Before i-vector extraction, the feature-domain channel compensation such as fLFA [9] will be applied to the acoustic features for better performance. Low-dimensional new i-vectors after PCA or LDA also make it easy for various classifiers with avoiding the curse of dimensionality.

In this paper, we model with two classifiers: cosine distance scoring (CDS) and support vector machines (SVM), which are both widely used in i-vector based SRE system [4] with low complexity.

The remainder of this paper is organized as follows: In section 2, the proposed combination method in i-vector level is introduced and then two classifiers CDS and SVM are briefly described in section 3. Experimental setup is present and the results are showed in section 4. Finally, we summarize the experimental results and give conclusion in section 5.

2. Combination method in i-vector level

2.1. Acoustic feature extraction

In this work, we used two acoustic complementary features extracted from the basic features. The first one is the 56-dimension shifted delta cepstral (SDC) derived from the 13-dimension perceptual linear predictive (PLP) feature. Then the

SDC coefficients concatenated with a popular 7-1-3-7 scheme are obtained.

The second one is the 55-demension time frequency cepstrum (TFC) [8] from the 13-dimension Mel-frequency cepstral coefficients (MFCC). This feature is obtained by performing a temporal discrete cosine transform (DCT) on the cepstrum matrix and selecting the transformed elements in a zigzag scan order.

Vocal tract length normalization (VTLN) and relative spectral (RASTA) filtering are applied during PLP and MFCC basic feature extraction. In addition, both of the two acoustic features are compensated by fLFA to provide a better performance.

2.2. Combination method in i-vector level

At first, we extract two kinds of i-vectors from both two acoustic features. The concept of i-vector are motivated by the JFA, in which both speaker and intersession subspaces are modeled separately, while i-vector method models all the important variability in the same low dimensional subspace named total variability space for using the useful information in channel subspace. Hence, the estimation of low rank rectangular total variability space is much more like the eigenvoice adaptation in JFA [9].

Like in i-vector based SRE system, we suppose the language-dependent and channel-dependent GMM supervector adapted from universal background model for a given utterance in i-vector based LRE can also be modeled as follows:

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

where m is the language-independent and channel-independent component of the mean supervector (usually from UBM mean), T is a matrix of bases spanning the subspace covering both language- and session-specific variability in the super-vector space, and ω is a standard normally distributed latent variable. For each utterance, the final i-vector is the maximum a posteriori (MAP) point estimate of the latent variable ω. More about i-vector extraction procedure are detailed in [9]. LDA and within class covariance normalization (WCCN) [4] are applied to the i-vector based LRE system.

After i-vectors are extracted, we firstly simply concatenate the multiple i-vectors to a new high-dimension i-vector. Then, in order to reduce the dimension and to remove the useless information, we then apply the unsupervised PCA, supervised LDA to the concatenated i-vector, and evaluate the performance respectively.

3. Classifier

3.1. Cosine distance scoring

In i-vector based speaker recognition system, the cosine distance scoring [4] has been proved the fastest and most efficient method, which directly uses the value of the cosine kernel between the target language i-vector and the test i-vector as a decision score. Following this way, we apply this modeling and scoring method in i-vector based LRE system as follow:

$$K(\omega_{lang}, \omega_{test}) = \frac{< \omega_{lang}, \omega_{test} >}{\left\| \omega_{lang} \right\| \bullet \left\| \omega_{test} \right\|} \tag{2}$$

In (2), the ω_{lang} can be obtained by lots of train segments and ω_{test} be obtained by a single segments. The value of this kernel is directly used as the final scoring. By CDS, no target language enrollment is required, so it can make the modeling and scoring faster and less complex than other modeling methods.

3.2. Support vector machine

Support vector machine is a powerful supervised binary classifier that has been efficiently adopted in speaker recognition and language recognition [2]. The target of this classifier is to model the decision boundary between two classes as a separating hyper plane from a set of supervised data examples defined by $X = \{(x_1, y_1), (x_1, y_1), ..., (x_N, y_N)\}$. Through labeling the positive examples $(y_i = +1)$ and the negative examples $(y_i = -1)$, the linear separating hyper plane can be obtained by solving the function as follow:

$$\begin{aligned} &F : \mathbb{R}^N \to \mathbb{R} \\ &x \to f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \end{aligned} \tag{3}$$

Where x is an input vector and $(\alpha_i, b; i = 1 \cdots N)$ are the SVM parameters obtained during the training. The cosine kernel $K(\cdot, \cdot)$ that we adopted in this work is as the same as the (2).

4. Experimental setup

4.1. Experimental data

The training data used in our experiments include two classes: conversational telephone speech data (CTS) and broadcast news data (BN). The CTS dataset includes the data from multiple corpora such as the OGI, CallFriend, CallHome, and OHSU. The BN dataset includes the data from VOAs supplied by NIST or downloaded from the Internet. All these data are pooled together and selected randomly to be the training corpus. The evaluation data come from NIST LRE09 dataset [10], which contains 23 target languages and three duration conditions of 3s, 10s and 30s.

In our experiments, all data in train set are used to train the 1024-mixture UBM and the dimension of total variability space was set to 400 as in [4]. After processing of LDA+WCCN, the dimension of raw i-vector can reduce to 200 also as in [4]. Our experimental results show in closed-set pooled error equal rate (EER) without backend processing. At the end, we compare out combination method with the LLR fusion in score level.

4.2. Evaluation of combination in i-vector level

4.2.1. Performance of i-vector based LRE baseline systems

We first evaluate the performance of two kinds of baseline systems, respectively by CDS and by SVM. Each kind of baseline system also includes two sub systems, which are using i-vector derived from SDC and using i-vector derived from TFC respectively. The results in Table 1 and Table 2 show that using TFC with CDS classifier can provide the best performance in all four baseline systems.

Table 1. The performance (in EER) of two i-vector based LRE baseline systems by CDS

EER (%)	i-vec (SDC)	i-vec (TFC)
30 s	4.11	**4.00**
10 s	8.58	**8.57**
3 s	18.80	**18.80**

Table 2. The performance (in EER) of two i-vector based LRE baseline systems by SVM

EER (%)	i-vec (SDC)	i-vec (TFC)
30 s	5.57	**5.30**
10 s	11.30	**9.84**
3 s	23.07	**19.92**

4.2.2. Performance of i-vector simply concatenate method

Next, we evaluate the proposed i-vector simple concatenation method by both CDS and SVM classifiers. The results are shown in Table 3 and Table 4.

With the comparison of the results showed in Table 2 and Table 3, we can see that no matter using which kind of classifiers, the simple concatenation of two i-vectors can always provide the better performance than baselines, respectively. In our experiments, the two raw i-vectors are 200-dimension respectively, and the concatenated i-vector is 400-dimension. It is shown that the CDS classier performs much better than the SVM classifier. This result is consistent with the result in [4].

Table 3. The performance of i-vector concatenation by CDS

EER (%)	i-vec (SDC)	i-vec (TFC)	i-vec (concatenation)
30 s	4.11	4.00	3.62
10 s	8.58	8.57	7.66
3 s	18.80	18.80	17.30

Table 4. The performance of i-vector concatenation by SVM

EER (%)	i-vec (SDC)	i-vec (TFC)	i-vec (concatenation)
30 s	5.57	5.30	**4.58**
10 s	11.30	9.84	**9.29**
3 s	23.07	19.92	**19.90**

4.2.3. Performance of i-vector combination using PCA

We evaluate the performance of i-vector combination after using the unsupervised PCA to reduce the dimension to 260 with accounting for 95% of the variance by both two CDS and SVM. The results are shown in Table 5 and Table 6. We can see that PCA not only reduce the dimensionality of combined i-vectors, but also improve the performance slightly, especially for CDS classifier. The reason for this may be that PCA can make the language i-vectors become more discriminative.

Table 5. The performance comparison of combination method before and after using PCA by CDS

EER (%)	i-vec (concatenation)	i-vec (using PCA)
30 s	3.62	**3.32**
10 s	7.66	**7.45**
3 s	17.30	**17.29**

Table 6. The performance comparison of combination method before and after using PCA by SVM

EER (%)	i-vec (concatenation)	i-vec (using PCA)
30 s	4.58	**4.58**
10 s	9.29	**9.29**
3 s	19.90	**19.88**

4.2.4. Performance of i-vector combination using LDA

By using the supervised LDA to reduce the concatenated dimension, the raw 400 dimensions can reduce to 22 dimensions. The results are present in Table 7 and Table 8. We can see that the good performances of both two classifiers are still keeping. And the CDS classifier can provide the better performance.

Table 7. The performance comparison of combination method before and after using LDA by CDS

EER (%)	i-vec (concatenation)	i-vec (using LDA)
30s	3.62	**3.11**
10s	7.66	**6.83**
3s	17.30	**16.50**

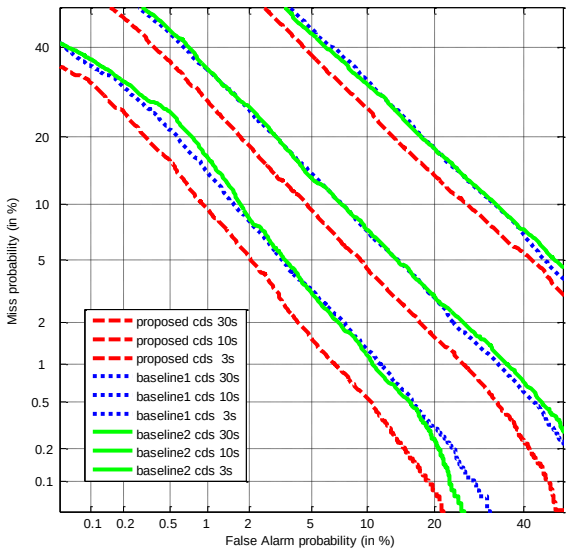

Figure.1. Performance of baseline systems and improved best system by CDS

Figure.1 and Figure.2 show the DET curves of two kinds of baseline systems in Table 1, and Table 2, and the improved systems in column 3 in Table 7 and in column 3 in Table 8. It shows that performance of the best-performed i-vector based CDS system proposed in this paper can reduce 1% in EER

corresponding to the baseline i-vector systems for 30 s and 2.3% in EER for 10 s and 3 s.

Table 8. The performance comparison of combination method before and after using LDA by SVM

EER (%)	i-vec (concatenation)	i-vec (using LDA)
30s	4.58	**4.58**
10s	9.29	**9.29**
3s	19.90	**19.76**

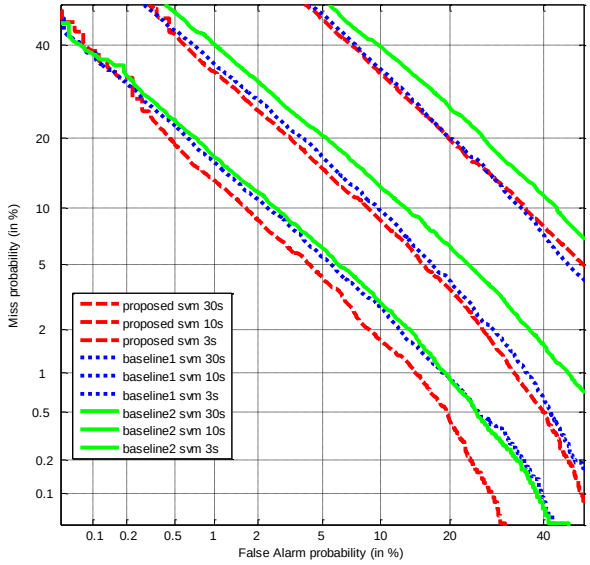

Figure.2. Performance of baseline systems and the best-improved system by SVM

4.2.5. Comparison with fusion in score level

For comparison with fusion in score level, we use the LLR to do the fusion by the scores in column 1 and 2 in Table 3 with the focal multiclass toolkit [10] and compare its performance with the scores in column 3 in Table 7 both after the score calibration. The results in Table 9 show that combination method in i-vector level with the score calibration can also provide the better performance than the fusion in score level. The reason for this may be that the combination in i-vector level can make use of the more discriminative information interweaving in the i-vector levels, while for the score-level fusion the information is already reduced to the single scores.

Table 9. The performance comparison of the best combination in i-vector level with the fusion in score level

EER (%)	fusion in score level	combination in i-vector level
30s	2.74	**2.63**
10s	6.29	**6.29**
3s	16.42	**16.37**

5. Conclusion

In this paper, we propose an effective complementary combination method in i-vector level for providing the better performance in LRE. PCA and LDA are used to reduce the high dimension and to remove the abundant information. Both CDS and SVM are applied to model the new combined i-vectors. The experimental results in NIST LRE2009 dataset show that the proposed complementary combination method in i-vector level can offer the better performance than fusion in score level. The performance of best system proposed in this paper can reduce 1% in EER than the relative baseline systems for 30 s duration and 2.3% in EER for 10 s and 3 s.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 60931160443, No. 61005019), by National High Technology Research and Development Program of China (No. 2008AA040201) and by National Science and Technology Pillar Program of China (No. 2009BAH41B01).

7. References

[1] L. Burget, P. Matejka, and J. Cernocky, "Discriminative training techniques for acoustic language identification," *in Proc. ICASSP*. vol.1. pp. 209-212, May 2006.

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

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transaction on Audio Speech and Language Processing*. vol. 15, no. 4, pp. 1435-1447, May 2007.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front end factor analysis for speaker verification", *IEEE Transaction on Audio, Speech and Language Processing*, vol. 19, no. 4, pp.788-798, May 2011.

[5] N. Dehak, P. Carrasquillo, D. Reynolds, R. Dehak, "Language recognition via ivectors and dimensionality reduction," *in Proc. Interspeech*, pp.857-860, Aug 2011.

[6] D. Martınez, O. Plchot, L. Burget, O. Glembek and P. Matejka "Language Recognition in i-vectors Space," *in Proc. Interspeech*, pp. 861-864, Aug 2011.

[7] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," *in Proc. IEEE Odyssey*, pp. 1-6, Jun. 2006.

[8] W.Q. Zhang, L. He, Y. Deng, J. Liu, and M. T. Johnson, "Time frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19. no. 2. pp. 266-272, Feb. 2011.

[9] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transaction on Speech Audio Processing*, vol. 13, no. 3, pp. 345-354, May. 2005.

[10] sites.google.com/site/nikobrummer/focalmulticlass.

[11] Kockmann Marcel, Ferrer Luciana, Burget Lukáš, Černocký, "ivector fusion of prosodic and cepstral features for speaker verification", *in Proc. Interspeech*, pp.265-268. Aug 2011