第45卷第2期 上海师范大学学报(自然科学版) Vol. 45, No. 2
2016年4月 Journal of Shanghai Normal University(Natural Sciences) Apr., 2016

DOI:10.3969/J. ISSN. 1000 - 5137. 2016. 02. 019

A novel L-vector representation and improved cosine distance kernel for Text-dependent Speaker Verification

LI Wei[1], YOU Hanxu[1], ZHU Jie[1], CHEN Ning[2]

(1. School of Electronic Information and Electical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; 2. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: A text-dependent i-vector extraction scheme and a lexicon-based binary vector (L-vector) representation are proposed to improve the performance of text-dependent speaker verification. An utterance used for enrollment or test is represented by these two vectors. An improved cosine distance kernel combining i-vector and L-vector is constructed to discriminate both speaker identity and lexical (or text) diversity with back-end support vector machine(SVM). Experiments are conducted on RSR 2015 Corpus part 1 and part 2. The results indicate that at most 30% improvement can be obtained compared with traditional i-vector baseline.

Key words: text-dependent speaker verification; i-vector; L-vector; cosine distance kernel

CLC number: TP 912.3 **Document code**: A **Article ID**: 1000-5137(2016)02-0243-05

1 Introduction

In recent years, i-vector based framework has demonstrated state-of-the-art performance in text-independent speaker verification[1]. Each utterance either for enrollment or test is projected onto a low rank total factor space, and is represented by a low dimensional identity vector termed i-vector. It is commonly thought that i-vector well captures speaker- and channel- dependent information in an utterance, also it represents a global adaptation in Gaussian Mixture Model (GMM) subspace. However its applicability has not been widely accepted in text-dependent speaker verification[2] mainly due to two reasons. Firstly, i-vector cannot explicitly represent the lexical information of an utterance. Secondly, since the duration of utterance is very short in text-dependent speaker verification, short-term speaker features, like Mel Frequency Cepstrum Coefficient (MFCC) or Perceptual Linear Predictive (PLP), can only activate a subset of total Gaussian components, hence it is not appropriate to globally adapt all the Gaussian components.

Received date: 2016-02-29

Foundation item: This work was supported by the National Natural Science Foundation of China (NSFC) under Grant (61271349, 61371147, 11433002), and Shanghai Jiao Tong University joint research fund for Biomedical Engineering under (YG2012ZD04).

Corresponding author: ZHU Jie, School of Electronic Information and Electical Engineering, Shanghai Jiao Tong University, No. 800, Dongchuan Rd., Shanghai 200240, China, E-mail: zhujie@ sjtu. edu. cn

To cope with these two shortcomings, firstly, we propose a text-dependent i-vector extraction scheme, only those Gaussian components with sufficient speaker frames are retained based on this scheme, and i-vector adaptation is performed based on this subset. Secondly, a lexicon-based binary vector termed L-vector is constructed to model the distribution of zero order Baum-Welch statistics, which can capture lexical information in an utterance. Finally, an improved cosine distance kernel is constructed, which combines i-vector and L-vector, to measure the diversity of both speaker identity and lexical (or text) content.

2 Text-dependent i-vector extraction

Given the speaker frame set of an utterance, we regard corresponding zero order Baum-Welch statistics N_c as a metric to measure how many frames are assigned to each Gaussian component, where c indexes each Gaussian component. According to[3], extremely short utterance (less than 10 s) leads to an imbalanced distribution of zero order Baum-Welch statistics, we can use 50% of total Gaussian components with highest N_c to capture more than 90% speaker frames. In text-dependent speaker verification, enrollment or test utterance is also very short, moreover, scarce N_c may lead to biased estimation of first order Baum-Welch statistics F_c[3], hence it is more appropriate to perform i-vector adaptation within a subset of total Gaussian components.

In order to select those Gaussian components with highest N_c, a threshold function is defined as:

$$S(c) = \begin{cases} 1 & N_c > \varepsilon \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

Where ε is an empirically tuned factor to adjust the number of Gaussian components to be retained. By this filter scheme, we can select a subset of Gaussian components with highest N_c. In real application, we usually pay more attention to the number of Gaussians in the subset, which we denote by R. The text-dependent i-vector extraction can be written as:

$$\tilde{w} = (I + \sum_{c=1}^{C} S(c) N_c(u) T_c^t \Sigma_c^{-1} \boldsymbol{T}_c) - 1 \sum_{c=1}^{C} T_c^t \Sigma_c^{-1} S(c) (F_c(u) - N_c(u) m_c). \tag{2}$$

where u denotes the utterance involved, I is the identity matrix as a prior, $\boldsymbol{T}_c$ is the sub-matrix of the c-th block of total factor matrix $\boldsymbol{T}$, $\boldsymbol{T}_c$ and m_c are the speaker-and text-independent covariance matrix and mean vector for c-th Gaussian component, C is the number of total Gaussian components. Compared to traditional i-vector extraction[4], the $S(c)$ filtering mechanism ensure that only those components representing lexical information of utterance involve in adaptation.

3 Lexicon-based L-vector

Although our improvedi-vector can be regarded as a text-dependent local representation in GMM space, it aims to discriminate speaker identity and cannot well discriminate lexical diversity. A lexicon-based binary vector termed L-vector is constructed for this purpose.

Utilizing the same $S(c)$ in (1), L-vector can be written as:

$$L = [S(1), S(2) \cdots S(C)] \in RC \times 1 \equiv [0_1, 1_2, 0_3, 1_4 \cdots 0_C]. \tag{3}$$

where the subscript indexes Gaussian component, the number of 1s in L is equal to R, the dimensionality of L-vector is equal to the number of total Gaussian components C. L-vector represents which Gaussian component is activated given a training utterance, and it encodes lexical information in utterance.

4 Improved cosine distance kernel

Given an enrollment utterance u_1 and a test utterance u_2, corresponding speaker models λ can be

represented as:

$$\lambda(u) = \{\tilde{w}(u), L(u)\}, u \in \{u_1, u_2\}. \tag{4}$$

To calculate the similarity between u_1 and u_2, the improved cosine distance kernel can be written as:

$$k(\lambda(u_1), \lambda(u_2)) = \frac{\tilde{w}(u_1) T \tilde{w}(u_2)}{\|\tilde{w}(u_1) T\| \|\tilde{w}(u_2)\|} * \frac{L(u_1) T L(u_2)}{\|L(u_1) T\| \|L(u_2)\|}, \tag{5}$$

where high score of $k(\cdot)$ comes from dual matches of both speaker identity in $\tilde{w}$ and lexical representation in L. Fig. 1 shows a simplified description of how $\tilde{w}$ and L collaborate to discriminate both speaker identity and lexical diversity.

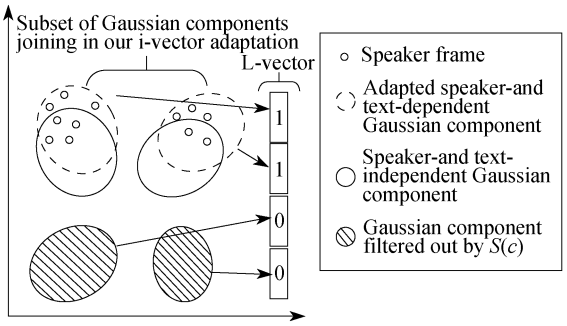

Figure 1　A brief description how $\tilde{w}$ and L collaborate.

5　Experiments and results

All experiments were carried out on part 1 and part 2 of the Robust Speaker Recognition 2015 (RSR 2015) corpus set[5-6], which is designed for text-dependent speaker recognition with scenario based on fixed pass-phrases (part 1) and fixed commands (part 2). It contains audio recordings from 300 people, which include 143 female and 157 male speakers that are between 17 to 42 years old, and the whole set is divided into background (bkg), development (dev) and evaluation (eval) subsets. Among the 300 people, 50 male and 47 female speakers are in the background set, 50/47 in the development set and 57/49 in the evaluation set.

Our experiments applied MFCC (19 order coefficients together with log energy) as short-term speaker feature, with speech/silence segmentation performed according to an energy-based voice activity detection (VAD). The length of Hamming window was 25ms with 10ms shift. The 20-dimensional feature vector was normalized by cepstral mean subtraction (CMS), 20 first order δ and 10 second order δ were appended, equal to a total dimension of 50.

512 order gender dependent universal background models (UBM) were trained with bkg corpus set. Gender dependent total factor matrixes with rank of 300 were trained with the mixture of bkg and dev corpus sets. In the back-end support vector machine (SVM) classification system, the speaker models λ extracted from bkg corpus set were used as imposter models to train the SVM system. Linear discriminant analysis (LDA) was applied as channel compensation technique before SVM training. LDA was estimated with the mixture of bkg and dev corpus sets. In our experiments, the optimal LDA dimension is 260. The eval set was used to evaluate system performance. Evaluations on part 1 and part 2 were independent and corpus sets between part 1 and part 2 were not overlapped. Two types of trials, i. e. CLIENT-wrong (given that the test utterance is spoken by the target user with wrong pass-phrase) and IMP-true (given that the test utterance is spoken by an imposter with the correct pass-phrase) of the evaluations described in[6] were used in our experiment. As we have mentioned in the previous section, the only parameter has to be empirically tuned in our system is R, R ranges from 512 ~ 350.

Results were given in terms of equal error rate (EER) and decision cost function (DCF). Table 1 and 2 present the results of traditional text-independent i-vector baseline system and our lexicon-based text-dependent i-vector system.

Tabel 1 Performance on evaluation trials of RSR2015 (part 1, eval set)

Types of trails	empirical parameter	MALE TRIALS		FEMALE TRIALS	
		EER(%)	DCF	EER(%)	DCF
CLIENT-WRONG	Baseline	2.61	0.020	3.07	0.022
	$R=500$	2.61	0.020	3.01	0.022
	$R=450$	2.27	0.018	2.68	0.021
	$R=430$	1.79	0.016	2.15	0.018
	$R=400$	1.87	0.017	2.40	0.020
	$R=350$	2.68	0.021	3.11	0.023
IMP-TRUE	Baseline	3.22	0.024	3.11	0.027
	$R=500$	3.22	0.024	3.85	0.027
	$R=450$	2.91	0.022	3.27	0.025
	$R=430$	3.10	0.022	3.39	0.025
	$R=400$	3.33	0.025	3.88	0.028
	$R=350$	4.12	0.030	4.50	0.036

Table 2 Performance on evaluation trials of RSR2015 (part 2, eval set)

Types of trails	empirical parameter	MALE TRIALS		FEMALE TRIALS	
		EER(%)	DCF	EER(%)	DCF
CLIENT-WRONG	Baseline	3.77	0.027	4.51	0.032
	$R=500$	3.72	0.026	4.46	0.031
	$R=450$	2.91	0.021	3.80	0.027
	$R=430$	2.80	0.020	3.61	0.025
	$R=400$	3.05	0.022	3.86	0.027
	$R=350$	3.99	0.029	4.70	0.034
IMP-TRUE	Baseline	6.70	0.038	8.05	0.043
	$R=500$	6.43	0.036	7.89	0.041
	$R=450$	6.07	0.031	7.18	0.036
	$R=430$	6.16	0.033	7.60	0.039
	$R=400$	6.76	0.039	8.21	0.045
	$R=350$	7.46	0.043	9..02	0.049

The results in both Table 1 and Table 2 show that as the value of R decreases from 512 ~ 430 (for CLIENT-wrong) or 450 (for IMP-TRUE), the system gains a significant performance improvement. In the CLIENT-wrong trials, best improvement is obtained when R is set to 430, our lexicon-based text-dependent i-vector system achieves a relative improvement of 26% in part 1 and 28% in part 2 on male trials as well as 30% in part 1 and 20% in part 2 on female trials. In the IMP-TRUE trials, as the lexical contents of target speaker and imposter speaker are identical, our lexicon-based text-dependent i-vector system gains less significant improvement, best improvement is obtained when R is set to 450, which achieves a relative improvement of 9.7% in part 1 and 9.5% in part 2 on male trials as well as 15% in part 1 and 10.9% in part 2 on female trials. In real application, setting R to 430 can obtain a global optimal performance in our text-dependent i-vector system.

6 Conclusion

We have proposed a lexicon-based local representation algorithm for text-dependenti-vector speaker verifi-

cation system. A subset of total Gaussian components is selected, which is most relevant to lexicon information. Text-dependent i-vector for either enrollment utterance or test utterance is extracted based on this subset. Moreover, a lexicon-based L-vector is constructed to discriminate lexical diversity. An improved cosine kernel is designed to measure the similarity of both speaker identity and lexical content between two utterances. Experimental results show that at most 30% improvement in EER can be obtained compared to traditional text-independent i-vector system. Given that our system now still highly depend on the empirical value R, our future work will focus on adaptive approach for tuning R automatically from speaker data.

References:

[1] Dehak N, Kenny P, Dehak R, et al. Front-end factor analysis for speaker verification [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(4): 788 – 798.

[2] Aronowitz H. Text dependent speaker verification using a small development set[C]//Odyssey. The Speaker and Language Recognition Workshop. ISCA: Singapore, 2012.

[3] Li W, Fu T F, Zhu J, et al. Sparsity Analysis and Compensation fori-Vector Based Speaker Verification[M]//Ronzhin A, Potapova R, Fakotakis N. Speech and Computer. Berlin: Springer International Publishing, 2015: 381 – 388.

[4] Kenny P, Boulianne G, Dumouchel P. Eigenvoice modeling with sparse training data [J]. Speech and Audio Processing, IEEE Transactions on, 2005, 13(3): 345 – 354.

[5] Larcher A, Lee K A, Ma B, et al. Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances[C]//IEEE. Acoustics Speech and Signal Processing (ICASSP) 2013 IEEE International Conference on. IEEE, Vancouer, 2013: 7673 – 7677.

[6] Larcher A, Lee K A, Ma B, et al. RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases [C]//Institute for Information Research. Interspeech. IZR: Singapore, 2012.

一种应用于文本相关说话人确认的 L-向量表示和改进的余弦距离核函数

李　为[1], 游寒旭[1], 朱　杰[1], 陈　宁[2]

(1. 上海交通大学 电子信息与电气工程学院, 上海 200240;
2. 华东理工大学 信息科学与工程学院, 上海 200237)

摘　要: 提出了一种用于文本相关说说话人确认技术的 i-向量提取方法和 L-向量表示. 一段用于注册或识别的语音可以用 i-向量和 L-向量联合表示. 同时提出了一种改进的用于支持向量机(SVM)后端分类的核函数,改进的核函数可以同时区分说话人身份的差异和文本内容的差异. 在 RSR 2015 语料集合 1 和集合 2 上验证系统的性能,实验结果显示改进的算法相对于传统的 i-向量系统的基线能提高至多 30% 的识别率.

关键词: 文本相关说话人识别; i-向量; L-向量; 余弦核函数

(责任编辑:包震宇)