Lexicon-Based Local Representation for Text-Dependent Speaker Verification

Hanxu YOU†(a), Wei LI†, Nonmembers, Lianqiang LI†, Student Member, and Jie ZHU†, Nonmember

SUMMARY 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. I-vector and L-vector are used to represent the utterances for enrollment and test. An improved cosine distance kernel is constructed by combining i-vector and L-vector together and is used to distinguish both speaker identity and lexical (or text) content, an improved cosine distance kernel is constructed by combining i-vector and L-vector.

The rest of this article is organized as follows. In Sect. 2, we propose a new text-dependent i-vector extraction scheme. In Sect. 3, L-vector is introduced. An improved cosine distance kernel is constructed in Sect. 4. Experiments are conducted and several results are presented and analyzed in Sect. 5. Finally, we draw our conclusions in Sect. 6.

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 seconds) 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 \( F_c \), hence it is more appropriate to perform i-vector adaptation within a subset of total Gaussian components. We use \( C \) to denote the number of total Gaussian components in an i-vector. Generally, for text independent tasks, bigger \( C \) always brings better performance speaker verification. In practice, to keep a good tradeoff between computational complexity and performance of speaker verification, the value of \( C \) is set to 512 by experience.

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

\[
S(c) = \begin{cases} 
1 & \text{if } N_c \geq \epsilon \\
0 & \text{Otherwise} 
\end{cases}
\]  

(1)

Where \( c \) denotes the index of Gaussian components, \( \epsilon \) 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:

\[
\hat{w} = \left( I + \sum_{c=1}^{C} S(c)N_c(\mathbf{a})^T\Sigma_c^{-1}T_c \right)^{-1}
\]
where \( u \) denotes the utterance involved, \( I \) is the identity matrix as a prior, \( T_c \) is the sub-matrix of the \( c \)-th block of total factor matrix \( T, \Sigma_c \) 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 key components representing lexical information of utterance involve in adaptation. We use \( \alpha = R/C \) to denote the ratio of the number of key components to the number of total components. As the value of \( C \) is chosen empirically, considering the lexical information of utterance, we try to find the most suitable \( \alpha \) for achieving performance improvement in text-dependent speaker verification tasks.

3. Lexicon-Based Binary Vector

As a text-dependent local representation in GMM space, our improved i-vector aims to distinguish speaker identity, but it does not seem to work on distinguishing lexical diversity. A lexicon-based binary vector termed L-vector is constructed to improve the performance. Utilizing the same \( S_c \) in (1), L-vector can be written as:

\[
L = [\mathcal{S}(1), \mathcal{S}(2), \ldots, \mathcal{S}(C)] \in \mathbb{R}^{C \times 1} \\
\equiv [0_1, 1_2, 0_3, 1_4, \ldots, 0_C] 
\]

(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 \( \lambda \) can be represented as:

\[
\lambda(u) = [\hat{w}(u), L(u)], \quad u \in \{u_1, u_2\} 
\]

(4)

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

\[
k(\lambda(u_1), \lambda(u_2)) = \frac{\hat{w}(u_1)^T \hat{w}(u_2)}{||\hat{w}(u_1)|| ||\hat{w}(u_2)||} \\
= \frac{L(u_1)^T L(u_2)}{||L(u_1)|| ||L(u_2)||} 
\]

(5)

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

5. Experiments and Results

We conducted experiments to study the influence of parameter \( \alpha \) on the performance of text-dependent speaker verification. All experiments were conducted 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 [7], 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 delta and 10 second order delta were appended, so that the total dimension is 50.

As we mentioned in Sect. 2, \( C = 512 \) order gender dependent universal background models (UBM) were trained with bkg corpus set. Gender dependent total factor matrices 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 \( \lambda \) 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 [5] were used in our experi-

![Fig. 1](image-url)
Table 1  Performance on evaluation trials of RSR2015 (part 1, eval set).

| RSR part 1 | Male trials | Female trials |
|------------|-------------|---------------|
|            | EER (%)  | DCF | EER (%) | DCF |
| 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 |
| $R = 400$ | 1.57 | 0.017 | 2.40 | 0.020 |
| $R = 350$ | 2.68 | 0.021 | 3.11 | 0.023 |
| IMP-wrong  | baseline | 3.22 | 0.024 | 3.85 | 0.027 |
| $R = 500$ | 3.22 | 0.024 | 3.81 | 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).

| RSR part 2 | Male trials | Female trials |
|------------|-------------|---------------|
|            | EER (%)  | DCF | EER (%) | DCF |
| baseline   | 3.77 | 0.027 | 4.51 | 0.032 |
| $R = 500$ | 3.77 | 0.026 | 4.46 | 0.030 |
| $R = 450$ | 2.91 | 0.020 | 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-wrong  | 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 | 7.00 | 0.039 |
| $R = 400$ | 6.76 | 0.039 | 8.21 | 0.045 |
| $R = 350$ | 7.46 | 0.043 | 9.02 | 0.049 |

6. Conclusion

We have proposed a lexicon-based local representation algorithm for text-dependent i-vector speaker verification system. The components that are the most relevant to lexicon information are selected from total Gaussian components and form a subset. Improved text-dependent i-vector for either enrollment or test is extracted based on the subset. Moreover, a lexicon-based L-vector is constructed to distinguish lexical diversity. The similarity of both speaker identity and lexical content between two utterances is measured by a designed cosine kernel. Given that $C = 512$ is adequate to capture both inter- and intra-speaker variabilities, $\alpha \approx 86\%$ is recommended for performance improvement and experimental results show that at most $30\%$ improvement in EER can be obtained compared to traditional text-independent i-vector system.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61271349, 61371147 and 11433002.

References

[1] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans. Audio Speech Language Process., vol.19, no.4, pp.788–790, 2011.

[2] H. Aronowitz, “Text dependent speaker verification using a small development set,” Odyssey 2012—The Speaker and Language Recognition Workshop, 2012.

[3] W. Li, T.F. Fu, J. Zhu, and N. Chen, “Sparsity analysis and compensation for i-vector based speaker verification,” Lecture Notes in Computer Science, vol.9319, pp.381–388, 2015.

[4] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” IEEE Trans. Speech Audio Process., vol.13, no.3, pp.345–354, 2005.

[5] A. Larcher, K.A. Lee, B. Ma, and H. Li, “The RSR2015: Database for text-dependent speaker verification using multiple pass-phrases,” INTERSPEECH 2012, 2012.

[6] A. Larcher, K.A. Lee, B. Ma, and H. Li, “Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances,” 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.7673–7677, 2013.

[7] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, “Speaker recognition by combining MFCC and phase information in noisy conditions,” IEICE Trans. Inf. & Syst., vol.E93-D, no.9, pp.2397–2406, Sept. 2010.