A New Score Normalization for Text-Independent Speaker Verification

Hongke Ning, Y. X. ZOU*, Xuyan Hu
ADSPLAB/ELIP, School of ECE, Peking University
Shenzhen 518055, China
zouyx@pkusz.edu.cn

***Abstract*—In iVector-based speaker verification system, the claimed speaker was verified if the similarity between the iVector of the tested utterance (iVector-ts) and the iVector of the claimed speaker (iVector-cs) is smaller than a fixed threshold. The commonly used method to measure the similarity between the iVector-ts and iVector-cs is the cosine similarity scoring method. To further improve the performance of the speaker verification system when the training data is insufficient, a new scoring method termed as ratio normalization (Rnorm) scoring method is proposed, where the similarity between iVector-ts and iVector-cs is normalized by the dissimilarity between the tested speaker model and the universal background model (UBM). Preliminary experimental results with Timit database and self-built database show that our proposed Rnorm scoring method is able to reduce the equal error rate (EER) of the iVector-based TIV speaker verification system compared with that of using conventional cosine similarity scoring method.**

Keywords—cosine similarity scoring, iVector-based speaker verification, TIMIT, ratio normalization

I. INTRODUCTION

Voiceprint is one of the unique biometric and has found wide applications including access control, providing forensic evidence, and user authentication in telephone banking, etc. Speaker verification aims at using voiceprint to verify the identity of the claimed speaker [1]. Essentially, speaker verification is a process to accept or reject the identity claim of a speaker by comparing a set of measurements of the tested speaker's utterances with a reference set of measurements of the utterance of the person whose identity is claimed. There are two categories of speaker verification systems. One is called text-independent verification (TIV) [1] [2] [3], the other is text-dependent verification (TDV) [1] [4]. For TIV speaker verification system, there are no constraints on the words which the speakers are allowed to use. Thus, the claimed speaker's training utterance and the tested speaker's utterance may have completely different content, and the verification system must take the phonetic mismatch into account. In this research, we only consider the iVector-based TIV system, which was firstly proposed by Reynolds in [2].

To make the presentation clear, the diagram of an automatic text-independent speaker verification system is illustrated in Fig. 1. Fig. 1.(a) shows the speaker enrollment process, where the iVector of each speaker is computed and stored as an iVector-cs. Specifically, the feature extraction module computes the *mel-frequency cepstral coefficients* (MFCCs) by the method proposed in [5]. MFCCs were firstly introduced in early 1980s for speech recognition and then adopted in speaker verification. The MFCCs of an utterance are denoted as $\mathbf{X} = \{ \mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_K \}$, where $\mathbf{x}_k$ is an M-by-1 feature vector indexed at discrete time $k \in [1, 2, ..., K]$, K is the number of speech frames.

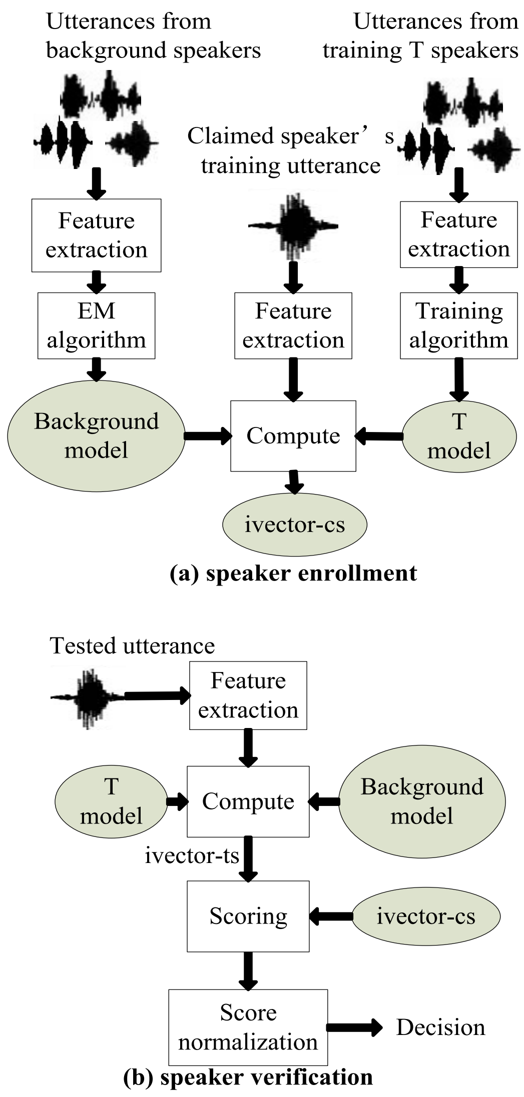

Fig. 1. Components of a typical automatic text-independent speaker verification system.

In Fig. 1.(a), the background model is firstly trained by *expectation maximization* (EM) training algorithm [6] taking MFCCs as the training data. The T model is trained exactly by the same process of learning the eigenvoice V matrix in [7] by taking MFCCs as the training data as well. Finally, the iVector-cs of the claimed speaker can be defined by taking the universal background model and T model [7]. Fig. 1.(b) presents the speaker verification process. The iVector-ts can be computed by the method proposed in [8] using the universal background model, T model and the MFCCs of the claimed speaker's utterance, which actually is of the same process as the computation of iVector-cs in Fig. 1(a). Then, the similarity of the iVector-ts and iVector-cs can be determined by the scoring module. Moreover, to achieve a better decision, a normalization module is added after the scoring module.

It is noted that there are a lot of researches have been conducted to evaluate the performance of the iVector-based TIV speaker verification system. For example, the impacts of background noise [9] [10], channel effect [11] [12], the duration of the utterance, and etc. To our knowledge, there are no research results evaluating the impact of the score normalization on the performance of the iVector-based TIV speaker verification system when the training data are insufficient. In this paper, we firstly evaluate the impact of the normalization and scoring on the performance of the iVector-based text-independent speaker verification system using TIMIT database. Aimed at reducing the EER of the system, a new score normalization method termed as *ratio normalization* (Rnorm) scoring method is proposed to normalize the similarity between iVector-ts and iVector-cs. The remainder of the paper is organized as follows. Section II describes the scoring method based on iVector-based TIV and the proposed Rnorm scoring method. Section III provides a description of four different conventional normalization scoring methods. Section IV details the experimental setup, experimental results and analysis of the results. The conclusions are given in section V.

II. Scoring Method

In this section, the scoring methods will be discussed in details. From Fig. 1.(b), it is easy to understand that the scoring method in the iVector-based TIV system is essentially to compute the similarity of the iVector-ts and iVector-cs. In the following, let's define scoring function as $score(\boldsymbol{\mu}, \mathbf{v})$. The popular cosine similarity (CS) function is given as follows:

$$score(\boldsymbol{\mu}, \mathbf{v}) = \frac{(\boldsymbol{\mu})^T \mathbf{v}}{\|\boldsymbol{\mu}\| \cdot \|\mathbf{v}\|} \tag{1}$$

where T is the transform operation. It is clear that the $score(\boldsymbol{\mu}, \mathbf{v})$ in Eq.(1) measures the similarity between vector $\boldsymbol{\mu}$ and $\mathbf{v}$ by their angle only, which does not consider the strength of the vectors.

A. Cosine Similarity Scoring Method

For iVector-based TIV speaker verification system, the commonly used scoring method is based on the CS function, which can be described as follows:

$$\Lambda_1(\mathbf{w}_{ts}, \mathbf{w}_{cs}) = score(\mathbf{w}_{ts}, \mathbf{w}_{cs}) \tag{2}$$

where $\mathbf{w}_{cs}$, $\mathbf{w}_{ts}$ are the N-by-1 iVector of the claimed and tested speaker's utterance, respectively. $score(\mathbf{w}_{ts}, \mathbf{w}_{cs})$ describes the cosine similarity between $\mathbf{w}_{ts}$ and $\mathbf{w}_{cs}$. The $\mathbf{w}_{cs}$ and$\mathbf{w}_{ts}$ are computed by the method proposed in [13]. The measure using Eq. (2) is termed as Cosine Similarity Scoring Method (CSS).

B. Proposed Ratio Normalization (Rnorm) Scoring Method

In this subsection, we proposed a *ratio normalization* (Rnorm) scoring method to normalize the similarity between the iVector-ts and iVector-cs.

To further improve the effective measure of the similarity between $\mathbf{w}_{ts}$ and $\mathbf{w}_{cs}$, we propose to normalize the similarity between $\mathbf{w}_{ts}$ and $\mathbf{w}_{cs}$ as follows:

$$\Lambda_2(\mathbf{w}_{ts}, \mathbf{w}_{cs}) = \frac{score(\mathbf{w}_{ts}, \mathbf{w}_{cs})}{score(\mathbf{w}_{ubm}, \mathbf{w}_{ts})} \tag{3}$$

where $score(\mathbf{w}_{ubm}, \mathbf{w}_{ts})$ presents the cosine similarity between $\mathbf{w}_{ubm}$ and $\mathbf{w}_{ts}$. From Eq.(3), we can see that if $score(\mathbf{w}_{ts}, \mathbf{w}_{cs})$ is given, $\Lambda_2(\mathbf{w}_{ts}, \mathbf{w}_{cs})$ is inversely proportional to $score(\mathbf{w}_{ubm}, \mathbf{w}_{ts})$. If $\mathbf{w}_{ts}$ is similar to $\mathbf{w}_{cs}$, in theory, $\mathbf{w}_{ts}$ is dissimilar to $\mathbf{w}_{ubm}$, $score(\mathbf{w}_{ubm}, \mathbf{w}_{ts})$ becomes smaller and $\Lambda_2(\mathbf{w}_{ts}, \mathbf{w}_{cs})$ goes larger. This is benefit for acceptance of $\mathbf{w}_{ts}$. In contrary, if $\mathbf{w}_{ts}$ is dissimilar to $\mathbf{w}_{cs}$, in theory, $\mathbf{w}_{ts}$ is similar to $\mathbf{w}_{ubm}$, $score(\mathbf{w}_{ubm}, \mathbf{w}_{ts})$ goes larger, and $\Lambda_2(\mathbf{w}_{ts}, \mathbf{w}_{cs})$ is attenuated. Moreover, it is noted that $\mathbf{w}_{ubm}$ is trained as a background iVector as shown in Fig. 1.(a). Comparing Eq.(2) and Eq.(3), we can see that Eq.(2) only measures the similarity between iVector-ts and iVector-cs, but Eq.(3) measures both the similarity between iVector-ts and iVector-cs, and the dissimilarity between the iVector-ts and iVector-cs. Therefore, we can conclude that $\Lambda_2(\mathbf{w}_{ts}, \mathbf{w}_{cs})$ given in Eq. (3) gives better measurement to accept or reject the tested utterance for iVector-based TIV speaker verification system and it is expected to give better performance. Considering the essential concept used to develop the scoring in ratio detector, in this paper, we term the method expressed in Eq.(3) as the *ratio normalization* (Rnorm) scoring method.

III. Conventional Scoring Method

The basic of the conventional normalization technique is to center the impostor score distribution by applying on each score generated by the speaker verification system. Since the study of Li and Porter [18], various kinds of score normalization techniques have been proposed in the literature. Four commonly used scoring methods by using different normalization approaches are briefly described in the following section.

A. Zero Normalization (Znorm) Scoring Method

The Znorm technique is directly derived from the work done in [14]. It has been massively used in the speaker verification systems in the middle of the nineties. The Znorm scoring method is defined as follows:

$$\Lambda_3(\mathbf{w}_{ts},\mathbf{w}_{cs}) = \frac{score(\mathbf{w}_{ts},\mathbf{w}_{cs}) - \mu_Z}{\sigma_Z} \tag{4}$$

where μ_Z and σ_Z are mean and standard variance, which are computed based on [14]. These two parameters are computed off-line in the speaker enrollment phase.

B. Test Normalization (Tnorm) Scoring Method

Still based on the estimate of mean and variance parameters to normalize impostor score distribution, *test normalization* (Tnorm), proposed in [3], and differs from Znorm by the use of impostor models instead of test speech signals. The Tnorm scoring method is defined as follows:

$$\Lambda_4(\mathbf{w}_{ts},\mathbf{w}_{cs}) = \frac{score(\mathbf{w}_{ts},\mathbf{w}_{cs}) - \mu_T}{\sigma_T} \tag{5}$$

where μ_T and σ_T are mean and standard variance as well, which are computed [3] These two parameters are computed on-line in the speaker verification phase.

C. ZTnorm Scoring Method

ZTnorm [14] applies Znorm to characterize the response of each speaker model to a variety of (impostor) test segments followed by Tnorm to compensate for the variations of the testing segments, such as duration and linguistic content. Hence, for ZTnorm scoring method, we have:

$$\Lambda_5(\mathbf{w}_{ts},\mathbf{w}_{cs}) = \frac{\dfrac{score(\mathbf{w}_{ts},\mathbf{w}_{cs}) - \mu_Z}{\sigma_Z} - \mu_T}{\sigma_T} \tag{6}$$

where μ_Z and σ_Z are computed as those in Znorm, μ_T and σ_T are computed as those in Tnorm.

D. TZnorm Scoring Method

TZnorm applies Tnorm to compensate for the variations of the testing segments, such as duration and linguistic content, followed by Znorm to characterize the response of each speaker model to a variety of (impostor) test segments.

For TZnorm scoring method, the score is computed as:

$$\Lambda_6(\mathbf{w}_{ts},\mathbf{w}_{cs}) = \frac{\dfrac{score(\mathbf{w}_{ts},\mathbf{w}_{cs}) - \mu_T}{\sigma_T} - \mu_Z}{\sigma_Z} \tag{7}$$

where μ_Z and σ_Z are computed as in Znorm, μ_T and σ_T are computed as in Tnorm.

IV. EXPERIMENT

A. Experiment Setup

Visual inspection of the DET curve [16] and equal error rate (EER) is commonly used evaluation tools in the speaker verification literature, which were proposed in the 1990s. The equal error rate (EER) is the standard scalar measure of the performance of a biometric verification system. In essence, it is the point on the DET curve where the false acceptance rate and the false rejection rate are equal. Moreover, NIST uses a detection cost function (DCF) as the primary evaluation metric to assess speaker verification performance. So we use EER and DCF as our evaluation metric in this paper.

The performance of speaker verification described above will be evaluated with the experimental setup as follows: 1) The iVector-based TIV speaker verification system is implemented and evaluated by ALIZE toolkit [17]. 2) For all evaluations, 64 speakers form the TIMIT speech corpus [3] are selected for speaker verification. The speakers are evenly balanced between the 8 different dialects and gender (i.e. 32 male and 32 female speakers with 4 male and 4 female speakers from each dialect region). The 10 utterances per speaker are divided into 5 utterances for training and 5 utterances for testing. 3) Our experiments operate on cepstral features, extracted using a 25 ms with Hamming window. 19 mel frequency cepstral coefficients together with log energy are calculated every 10 ms. Delta and double delta coefficients are then calculated to produce 60-dimensional feature vectors. We use gender dependent universal background models containing 512 Gaussians and 400 total factors defined by the total variability matrix, which is trained using 320 utterances. The decision score obtained with cosine similarity scoring are normalized using Znorm and Tnorm. We use 480 Znorm utterances. 4) For convenience and presentation clarity, we have the following abbreviations: CSS means the method given in Eq.(2); Rnorm-CSS, Znorm-CSS, Tnorm-CSS, ZTnorm-CSS and TZnorm-CSS stand for the scoring methods given in Eq.(3), Eq.(4), Eq.(5), Eq.(6) and Eq.(7), respectively, where $score(\mathbf{w}_{ts},\mathbf{w}_{cs})$, $score(\mathbf{w}_{ubm},\mathbf{w}_{ts})$ are both computed by Eq.(1).

TABLE I. EER AND DCF RESULTS USING TIMIT DATABASE

Algorithm	Female		Male	
	EER(%)	*DCF*	*EER(%)*	*DCF*
CSS	2.9688	0.0300	2.9639	0.0299
Znorm-CSS	2.9883	0.0302	2.9688	0.0300
Tnorm-CSS	2.9688	0.0300	2.9688	0.0300
ZTnorm-CSS	2.9883	0.0302	2.9688	0.0300
TZnorm-CSS	2.9688	0.0300	2.9688	0.0300
Rnorm-CSS	2.9688	0.0300	**2.9102**	**0.0294**

B. Experiment 1: Performance Comparison with CSS and Five Different Normalization Scoring Methods

This experiment is conducted to compare the performance of the iVector-based TIV system using different normalization scoring methods described in this paper. The results are listed in Table I. From the results, we can see that all methods have quite similar EER and DCF performance under the experimental conditions for female speakers. Our proposed Rnorm-CSS method achieves better results than other comparison methods in male trials. For example, the Rnorm-CSS method gives an EER of 2.9102, but CSS method gets 2.9639. However, the CSS and Rnorm-CSS method give the same EER for females. These results may tell the facts that non-speaker information (such as session and channel) affects the iVector magnitudes. The proposed Rnorm-CSS method is able to remove the impact of the magnitude and then it is able to improve the robustness of the system.

C. Experiment 2: Performance Comparison with Four Scoring Methods on ADSP_SV Database

This experiment is carried out to evaluate the performance of the iVector-based TIV speaker verification system using the proposed Rnorm-CSS method on ADSP_SV database, which was recorded in our laboratory. This evaluation uses 9 speakers from the ADSP_SV database, each speaker have 20 sessions in Chinese, and they are recorded by two different microphones. The microphones used to record training and testing utterances are different. The experimental results are given in Table II. It is clear to see that the proposed Rnorm-CSS method outperforms CSS method and Tnorm–CSS methods. Moreover, Tnorm–CSS method generally achieves better results than that of the CSS method and Znorm-CSS method. An explanation of this may be that the utterances for training μ_Z and σ_Z is different from the utterances for training iVector-cs, but this mismatch don't exist in training μ_T and σ_T.

TABLE II. EER AND DCF USING ADSP_SV DATABASE.

System EER(%) [DCF]	Mic1_Mic1	Mic1_Mic2	Mic2_Mic1	Mic2_Mic2
CSS	9.38 [0.0947]	9.63 [0.0972]	9.88 [0.0997]	9.63 [0.0947]
Znorm-CSS	9.63 [0.0972]	9.88 [0.0997]	9.88 [0.0997]	9.38 [0.0947]
Tnorm-CSS	**8.89 [0.0897]**	9.38 [0.0947]	9.63 [0.0972]	8.89 [0.0897]
Rnorm-CSS	**8.89 [0.0897]**	**9.37 [0.0935]**	**9.14 [0.0922]**	**8.4 [0.0847]**

V. CONCLUSIONS

In order to improve the performance of the iVector-based TIV speaker verification system, this paper proposed a new normalization scoring method where the similarity between the tested speaker's iVector (iVector-ts) and the claimed speaker's iVector (iVector-cs) is normalized by the dissimilarity between the tested speaker model and the UBM. Intensive experiments with TIMIT database and self-built database have been carried out to evaluate the performance under insufficient training data. Experimental results showed that the proposed Rnorm-CSS method decreases the EER for male from 2.9639% to 2.9102%. Moreover, according to the evaluation of the performance of different normalizations on self-built database, Rnorm-CSS method achieves the best results for cross-channel iVector-based TIV system.

ACKNOWLEDGMENT

This work is partially supported by National Natural Science Foundation of China (No: 61271309) and Shenzhen Science & Technology Research Program.

REFERENCES

[1] Reynolds, D.A, Speaker identification and verification using Gaussian mixture speaker models. Speech communication, 1995. 17(1): p. 91-108.

[2] Reynolds, D.A. and R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models., IEEE Transactions on Speech and Audio Processing, 1995. 3(1): p. 72-83.

[3] Auckenthaler, R, M. Carey, and H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems. Digital Signal Processing, 2000. 10(1): p. 42-54.

[4] Hébert, M., Text-dependent speaker recognition. Springer Handbook of Speech Processing, 2008: p. 743-762.

[5] Davis, S. and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 1980. 28(4): p. 357-366.

[6] Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, Speaker verification using adapted Gaussian mixture models. Digital signal processing, 2000. 10(1): p. 19-41.

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

[8] Sardana, S. and A. Garg, Binary Mask for Speech Enhancement. International journal of electronics and data communication, 2013. 2(1): P. 1-3.

[9] Lei, Y, L. Burget, and N. Scheffer. A noise robust i-vector extractor using vector taylor series for speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013.

[10] Ferrer L, McLaren M, Scheffer N, et al. A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation[J]. 2013.

[11] McLaren, M., et al. Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013.

[12] Kanagasundaram, A, et al, I-vector based speaker recognition using advanced channel compensation techniques. Computer Speech and Language, 2014. 28(1): p. 121-140.

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

[14] Li, K.-P. and J.E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. Acoustics, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1988. IEEE.

[15] Malegaonkar, A., et al. On the enhancement of speaker identification accuracy using weighted bilateral scoring. IEEE International Carnahan Conference on Security Technology (ICCST). 2008. IEEE.

[16] Martin, A, Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. The DET curve in assessment of detection task performance. In Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997), pp. 1895-1898.

[17] Bonastre J F, Wils F, Meignier S. ALIZE, a free toolkit for speaker recognition[C]//proc. ICASSP. 2005, 5: 737-740.

[18] Li K P, Porter J E. Normalizations and selection of speech segments for speaker recognition scoring[C]//, International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1988: 595-598.

[19] Barras C, Gauvain J L. Feature and score normalization for speaker verification of cellular data[C]// 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, IEEE.