International Journal of Advanced Technology and Engineering Exploration, Vol 7(62)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2019.650076

Research Article

Text independent voiceprint recognition model based on I-vector

Jing Zhang* **and Minfeng Yao**
School of Information Science and Technology Guangdong University of Foreign Studies Guangzhou City, China

Received: 20-November-2019; Revised: 10-January-2020; Accepted: 12-January-2020
©2020 Jing Zhang and Minfeng Yao. This is an open access article distributed under the Creative Commons Attribution (CC BY) License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract
The commonly used text independent Voiceprint recognition models are Gaussian Mixture Model (GMM) and GMM and general background model (GMM-UBM). In the equalization vector of the GMM model, both the speaker information and the channel information are included, which results in unstable performance of the recognition system of the GMM and GMM-UBM models. In addition, the recognition ability for cross channel is poor, moreover, both models are limited by the maximum likelihood criterion. So, they employ weak ability to distinguish categories. I-vector is also known as identity authentication vector and has been proposed on the basis of Gaussian super vector in recent years. The method uses one space instead of the two spaces, including the difference between the speakers and the difference between the channels, and it is known as the most cutting-edge speaker modeling technology available today. Therefore, this paper adopted i-vector framework as the speaker recognition model, and studied the main problems that need to be dealt with. The recognition effect of GMM-UBM model and i-vector model were investigated by experiment as well. Through comparison experiments, it is verified that the i-vector recognition model employs a lower error rate of the and is more efficient. In the recognition phase, to quickly recognize the speaker's identity only needs to record two seconds of speech, and the system recognition accuracy reaches 97%.

Keywords
Speaker recognition, Text-independent, I-vector, EER.

1.Introduction
Speaker Recognition is a technology for identifying people based on the Voiceprint characteristics contained in voice, and is a kind of biometric authentication technology. According to different recognition methods, Voiceprint recognition can be divided into three categories [1]. (1) Speaker discrimination is to distinguish who is talking from a given speaker set; (2) speaker confirmation is to judge whether someone is speaking or not; (3) speaker segmenting and clustering is to distinguish the switching time of different speakers based on the voice formed by multiple speakers.

According to different text requirements, the Voiceprint recognition task can be divided into two types: (1) text-independent Voiceprint recognition, which means that there is no specific requirement for the corpus in the model training process, and the training corpus and the test corpus can also be different; (2) text-related Voiceprint recognition, that is, the user gets the training model according to the specified text pronunciation during model training, and the test text and training text should be consistent.

For the speaker recognition, the choice and establishment of the model are the most important. A high-performance recognition system should be provided with different models to adapt to different applications. The researchers provide a series of pattern matching methods for various characteristics with further research. In the field of text-independent Voiceprint recognition, GMM and GMM-UBM models have become the dominant recognition methods, and achieved relatively better recognition effect. However, in the EMM's equalization vector, both speaker information and channel information are

*Author for correspondence

This work is supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No. 17YJCZH242), School scientific research project (17TS16).

included, which results in poor cross-channel speech recognition capability of the GMM and GMM-UBM models [2], and the signal interference factor cannot be overcome. At the same time, due to the limitation of the maximum likelihood criterion, the ability to distinguish the categories is weak [3]. Therefore, some people have proposed the application of factor analysis to the speaker field [4−6]. The joint factor analysis considers that the super-vector of the Gaussian model in the GMM-UBM system can be roughly divided into a vector feature related to the speaker itself and a linear superposition of vector features related to the channel and other changes. That is to say, the space of GMM is divided into Eigen space, channel space and residual space. In this way, if the features related to the speaker itself can be extracted and the channel-related features be removed, then the channel influence can be better overcome and the recognition be carried out. It turns out that this kind of thinking is correct. After the joint factor analysis, the performance of the system is obviously improved [7].

The traditional modeling process of joint factor analysis is mainly based on two different spaces: the speaker space defined by the intrinsic speech space matrix, and the channel space defined by the intrinsic channel space matrix. Inspired by the theory of joint factor analysis, Dehak proposed to extract a more compact vector called the I-Vector from the GMM mean super-vector. The I-vector method uses one space instead of the two spaces. This new space can be a global difference space, which contains both the difference between the speakers and the difference between the channels.

Therefore, the I-Vector modeling process does not strictly distinguish the influence of the speaker and the influence of the channel in the GMM super-vector. It is called the most effective speaker modeling technique [8−10].

However, for i-vector model, there are many factors influencing the recognition performance, such as Gaussian mixture, LDA dimension, male to female ratio in the training set [11−12] and test voice duration. We have studied these factors and gave experimental results. Through experiments, the key factors to improve the performance of i-vector Voiceprint recognition model are given, and it is proved that the model can greatly improve the accuracy and real-time performance of Voiceprint recognition, which is very helpful in the practical application of the system.

2.The principle of i-vector

The basic idea of the i-vector based speaker recognition system is that assume both the speaker information and channel information are contained in the GMM's high-dimensional mean super-vector space, in this super-vector space, by training the Total Variability (TV) space which contains speaker information and the channel difference to decompose super-vector S of each speaker's speech data into Equation (1) [13].

$$S = m + Tw\ S = m + T\omega\ S = m + T\omega S = m + T\omega\ S = m + T\omega \quad (1)$$

Where S denotes the super-vector associated with the speaker and the channel; m denotes the super-vector independent of the speaker and the channel; the subspace matrix T of overall variation completes the mapping from the high dimensional space to the low dimensional space, thereby it makes the vector after dimensionality reduction is more conducive to further classifying and recognizing; ω represents the vector associated with the speaker channel, which is a full-variable spatial difference factor containing speaker information and channel information [14].

The framework of a speaker recognition system based i-vector is shown in *Figure 1*. The main problems that need to be processed are finding the full variable space T, and extracting the i-vector, and channel compensation as well as cosine distance scoring [15].

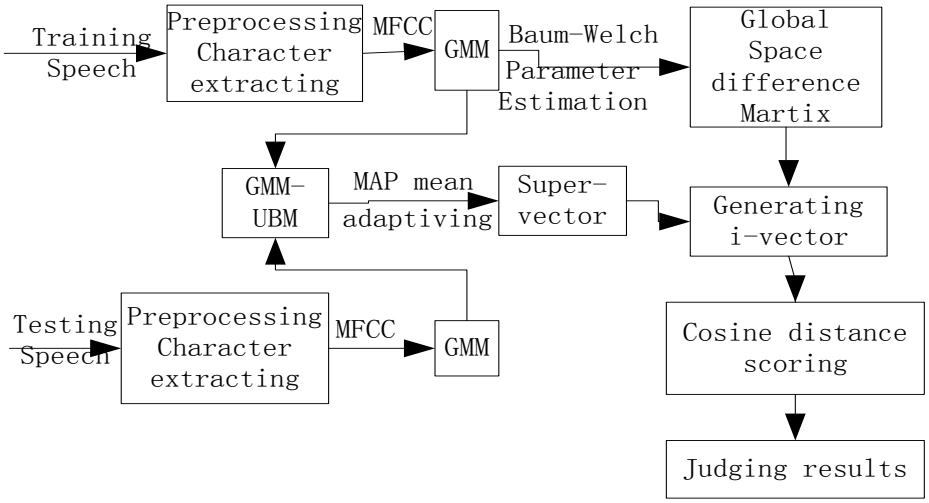

Figure1 I-vector recognition system

A. Full variable space matrix

For the k[th] segment speech of a given speaker A, the Baum-Welch statistic can be expressed as Equations 2, 3 and 4[16].

$$N_{j,k}(a) = \sum p(_j|x_t, \lambda) \quad (2)$$

$$\tilde{F}_{j,k}(a) = \sum p(j|x_t, \lambda)(x_t - m_j) \quad (3)$$

$$\tilde{S}_{j,k}(a) = diag[\sum p(j|x_t, \lambda)(x_t - m_j)(x_t - m_j)] \quad (4)$$

In the formula (2), $p(x_t, \lambda)$ is the posterior probability that the Gaussian mixture component j produces x_t in the UBM model, and $p(x_t, \lambda)$ can be expressed as (5).

$$p(x_t, \lambda) = \frac{\varepsilon_j p(x_t, \lambda)}{\sum_{j=1}^{M} \varepsilon_j p(x_t, \lambda)} \quad (5)$$

After obtaining the Baum-Welch statistic of all the speeches in the training inventory according to the above method, and to estimate the matrix T by 5 to 8 iterations using the EM algorithm, the T is considered to be converged, then generating a random initial value. The EM algorithm is as follows [17].

(1) Step E: Calculate the posterior distribution of ε, and obtain the expectation and correlation matrix of ε as shown in Equations 6, 7 and 8.

$$l^{-1}(a) = I + T^T \sum^{-1} N_k(a) T \quad (6)$$

$$E\left[\omega_{a,k}\right] = l^{-1}(a) T^T \sum^{-1} \tilde{F}_k(a) \quad (7)$$

$$E\left[\omega_{a,k}\omega^T{}_{a,k}\right] = E\left[\omega_{a,k}\right] E\left[\omega^T{}_{a,k}\right] + l^{-1}(a) \quad (8)$$

In Equation 6, $N_k(a)$ is a matrix takes the blocks matrix $N_{j,k}(a)I$ as diagonal and $\tilde{F}_k(a)$ in equation (7) is the high dimensional vector spliced by $\tilde{F}_{j,k}(a)$ Vectors. The is diagonal matrix whose diagonal elements are the covariance matrix of each order of the UBM.

(2) M step: To update the matrix T and maximize the likelihood function, as in Equation 9 and 10.

$$\varnothing_j = \sum_a \sum_k N_{j,k} E[\omega_{a,k}\omega^T{}_{a,k}] \quad (9)$$

$$\tau = \sum_a \sum_k \tilde{F}_k(a) E[\omega^T{}_{a,k}] \quad (10)$$

For each Gaussian mixture component m = 1, 2, ..., M and the one-dimensional f = 1, 2, ..., p of the characteristic parameter, let $i = (j - 1) \times p + f$, T_i represents the i-th line of T , and τ_i represents the i-throw of τ, then T can be updated to Equation 11.

$$T_i \varnothing_j = \tau_i, (i = 1, \ldots P) \quad (11)$$

In the Equation 11, P is a dimension of a characteristic parameter. The matrix T can be obtained by continuously iterating according to the E and M steps until it gets convergence.

B. Extracting i-vector vectors

After obtaining the zero-order and first-order Baum-Welch statistic of the speech, the corresponding i-

vector vector can be obtained by the following calculation, as in Equation 12.

$$E\left(\omega_{a,k}\right) = I^{-1}(a)\, T^{T} \sum^{-1} \tilde{F}_{k}(a) \tag{12}$$

C. LDA channel compensation algorithm

Channel compensation is for the channel information and the speaker's personal information in the whole variable space. By reducing the difference caused by the channel information as much as possible, and amplifying the speaker's personality information to achieve the purpose of enhancing the discrimination effect. Because i-vector uses the full variable space, it cannot be avoided that the system's distinguish ability will be affected by channel information. Therefore, channel compensation is needed to weaken the channel influence [18,19].

Linear discriminant analysis (LDA) is a dimensionality reduction technique widely used in the field of recognition. By maximizing the ratio of inter-class to intra-class dispersion and finding a set of orthogonal bases to project the original features to the space defined by base coordinates, then the effect of distinguishing ability of the characteristic can be improved.

Given a training set containing A speakers, and each speaker has a K_a segments of speech, i.e. K_a i-vectors, all the i-vectors constitute a training sample set. The criterion function of LDA is as shown in Equation 13.

$$D(\varphi) = \frac{\varphi^{T} S_{b}^{\varphi}}{\varphi^{T} S_{w}^{\varphi}} \tag{13}$$

In Equation 13, D (φ) is the Rayleigh coefficient, and φ is the optimal projected coordinate direction. The formula for calculating the inter-class covariance matrix S_b and the intra-class covariance matrix S_w is given by Equations 14 and 15.

$$S_b = \sum_{a=1}^{A} (\theta_a - \theta)(\theta_a - \theta)^{T} \tag{14}$$

$$S_w = \sum_{a=1}^{A} \frac{1}{K_a} (\gamma_i^a - \theta_a)(\gamma_i^a - \theta_a)^{T} \tag{15}$$

Where γ_i is the speaker's i-vector and θa is the mean of the a-th speaker's training sample. The Lagrange function is used to find the optimal solution of D(φ). After simplification obtains as shown in Equation 16.

$$S_b\varphi = \lambda S_\varphi \varphi \Rightarrow S_w^{-1} S_b^{v} = \lambda\varphi \tag{16}$$

D. Cosine distance scoring

Cosine distance scoring (CDS) is a symmetric kernel function classifier whose target vector and test vector exchange have no effect on the result [20]. The CDS calculates the cosine distance between the i-vector of the test speech and the i-vector of each speaker model, and compares it with the threshold to obtain the final decision result. The CDS calculation formula is as shown in equation (17).

$$Score = \frac{\langle \omega_{Sp}, \omega_{Ts} \rangle}{\| \omega_{Sp} \| \| \omega_{Ts} \|} \tag{17}$$

Where ω_{Sp} is the i-vector of speaker model, and ωT s is the i-vector of test speech model.

3.Speaker model experiment based on i-vector

A. Performance evaluation of speaker system

False rejection rate (FRR) is the probability that the voice from the real user is rejected by the system. The lower the FRR, the easier it is to accept the speaker. False acceptance rate (FAR) is the probability that the voice from the impostor is accepted by the system. The lower the FAR, the better the security of the system.

The detection error trade-offs curve (DET) is a curve in which FAR is the horizontal axis and FRR is the vertical axis, as shown in *Figure 2*.

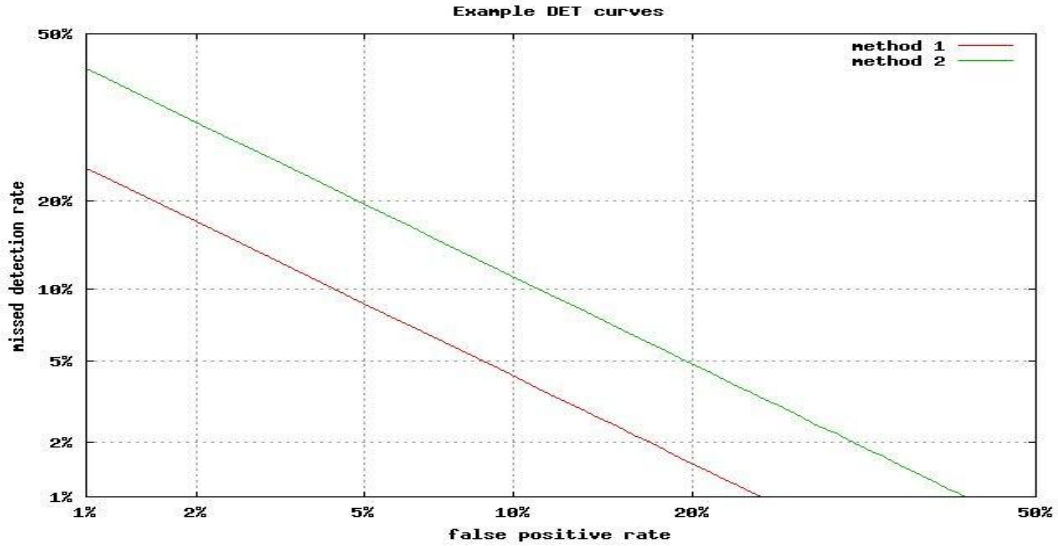

Figure 2 Detection error trade-off curve

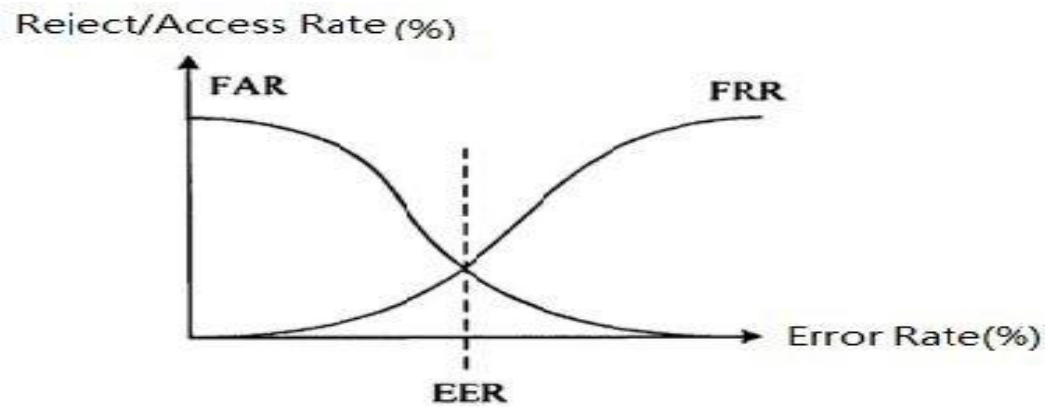

Figure 3 Relationship between threshold and error rate

The relationship between threshold and error rate is shown in *Figure 3*. It can be seen from *Figure 3* that no matter how the threshold is selected, neither FRR nor FAR can be reduced at the same time. Therefore, it is necessary to find an equal error rate (EER) point of FRR and FAR to select the decision threshold of the system. The EER of the speaker recognition system represents the average error rate and can be used to describe the average performance of the speaker recognition system.

B. The effect of male to female ratios on the recognition performance of GMM-UBM and i-vector systems

In this experiment, the MFCC is used as the characteristic parameter to explore the influence of different male-female mixing ratios on the performance of the GMM-UBM model and the i-vector based recognition system in the event that the training duration, the total number of training sets and the Gaussian mixture are unchanged.

The speeches used in this experiment were from TIMIT, which contains a total of 6,300 speech data from 630 people. A total of 4,000 speeches of 400 male and female students in the corpus were selected for UBM training. The proportion of males in different experiments was respectively 0%, 20%, 40%, 60%, 80%, and 100%. The test set selected 100 people for men and women, half of which were male and female. The DET obtained in each set of experiments is shown in Figure *4*. *Figure 5* is the statistics results, and *Table 1* was obtained after trimming the experimental data.

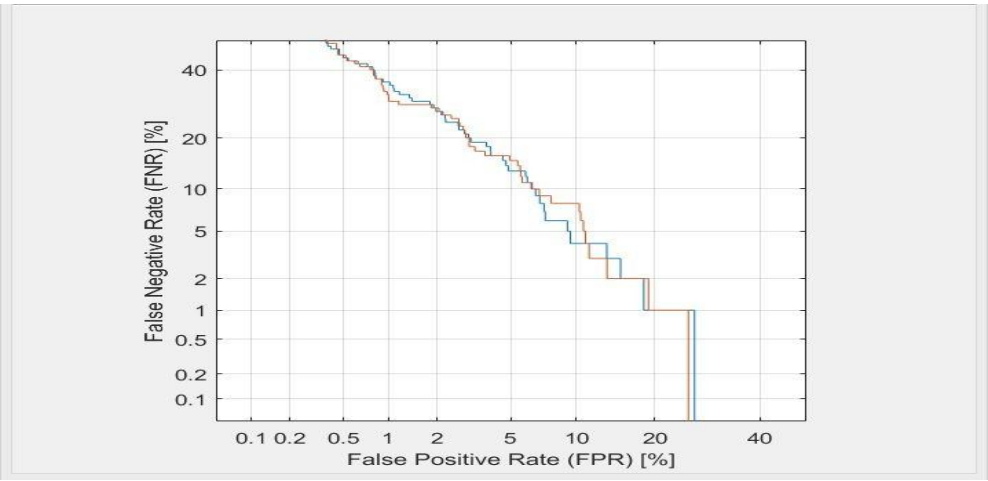

Figure 4 Experimental DET chart

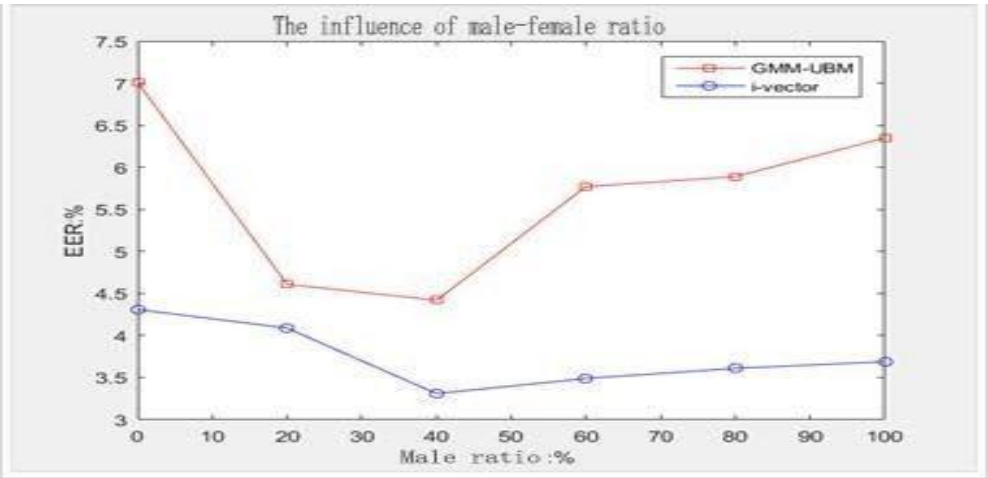

Figure 5 The effect of male-female ratio on system recognition rate

Table 1 The effect of male-female ratio on system identification rate

Male speech	GMM-UBM EER	i-vector EER
0%	7.01%	4.31%
20%	4.61%	4.09%
40%	4.42%	3.31%
60%	5.77%	3.49%
80%	5.89%	3.61%
100%	6.35%	3.69%

It can be seen from the experimental results that the male-female ratio in the training set will have a greater impact on the performance of the recognition system, so the male-female ratio should be considered in the training phase. When the male ratio is 40%, the recognition rate of the two recognition systems is the highest, which means that the corpus selected as training set with 4:6 male-female is the best. At the same time, the overall EER of i-vector is lower than that of GMM-UBM, especially when males' ratio is relatively large, the recognition of i-vector is more accurate than that of GMM-UBM.

C. Influence of gaussian mixture on recognition performance of GMM-UBM system and i-vector system

The purpose of this experiment is to investigate the influence of different Gaussian mixture on the recognition rate of GMM-UBM recognition system and i-vector recognition system. UBM training was conducted by using 530 people in the TIMIT corpus as a training set. 100 people were used as test sets, and the ratio of male to female in the test set was 40% and 60%. Under the same training set and test set conditions, to study the influence of Gaussian mixture on the performance of the recognition system by changing the Gaussian mixture degree for two different recognition systems. The obtained result data chart is shown in *Figure 6*, and the experimental data being sorted is shown in *Table 2*.

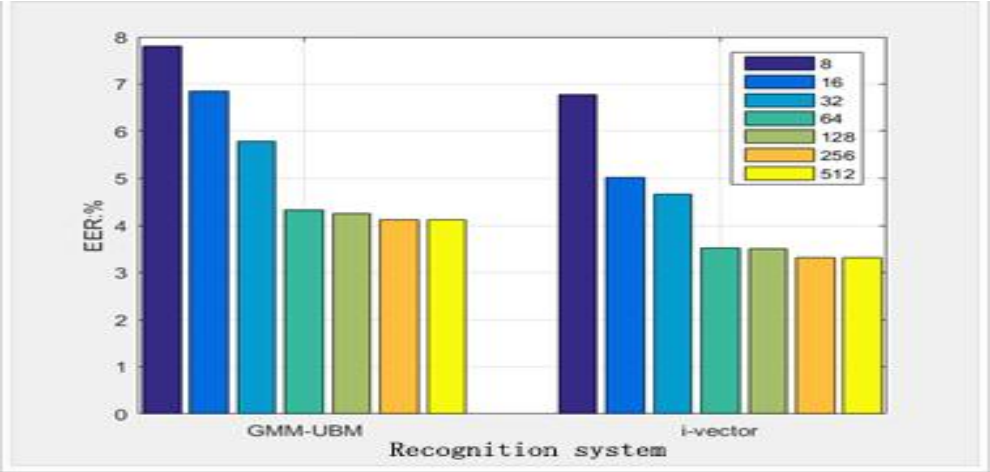

Figure 6 The influence of the mixture degree on the recognition rate

Table 2 The influence of Gaussian mixture degree on recognition rate

Gaussian mixture	EER of GMM-UBM	EER of i-vector
8	7.81%	6.78%
16	6.85%	5.01%
32	5.78%	4.66%
64	4.32%	3.53%
128	4.26%	3.51%
256	4.12%	3.32%
512	4.12%	3.31%

Although the recognition effect will increase with the increase of Gaussian mixture in theory, more factors need to be considered in the system implementation. The higher the Gaussian mixture, the larger the calculation, the greater the requirements on the equipment, and the more time consumption. Therefore, when choosing Gaussian mixture, the accuracy and calculation time should be comprehensively-considered.

The experimental results show that with the increase of Gaussian mixture, the error rate of the system is gradually reduced, which indicates that the increase of Gaussian mixture does improve the recognition rate. As the Gaussian mixture increases, the error rate decreases gradually, but when the mixture is mixed to 256, the error rate decreases very slowly. In order to avoid the computer's calculation amount too large, the mixing degree adopted in this experiment is 256.

D. Influence of LDA dimension on recognition performance of i-vector system

The purpose of this experiment is to investigate the influence of dimensionality reduction for LDA on the recognition performance of i-vectors. UBM training was conducted by using 530 people in the TIMIT corpus as a training set. 100 people were used as test sets, and the ratio of male to female in the test set was 40% and 60%. According to the data of Experiment 2, the Gaussian mixture is 256. The original dimension of i-vector is 200. *Figure 7* and *Table 3* show the reduced dimensions number of LDA and its influence on the recognition rate of the system.

The experimental results show that the error rate of the i-vector recognition system will decrease with the LDA dimension decreases, but when the LDA dimension decreases to a certain value, the system EER increases. According to the data in chart, for the system the lowest error rate is and the best recognition rate can be obtained when Dim is 90, so the LDA dimension selected by this experimental system is 90.

This experiment mainly investigated the use effect of the Android product in this project in the actual environment. Thereunto, the number of experimental people is 10 (5 males and 5 females), and the environment is the roadside and canteen. The experiment is mainly divided into the following three comparison modules.

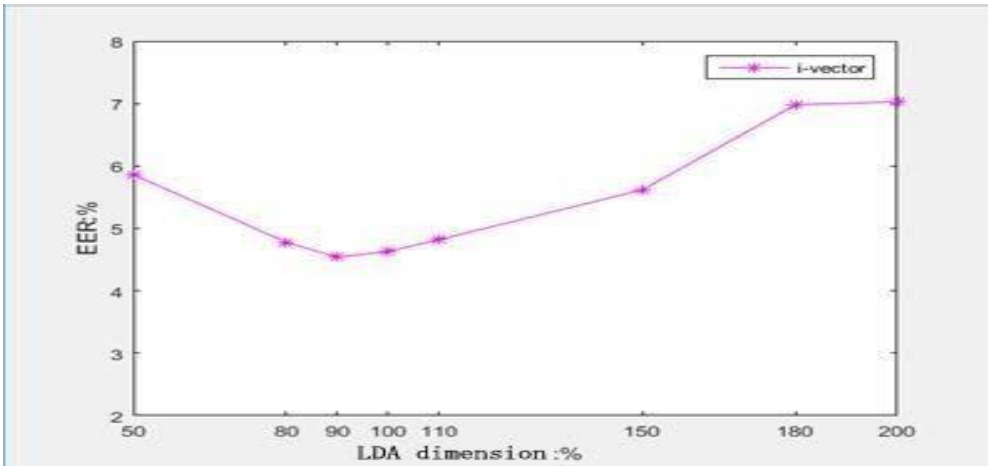

Figure 7 Result chart of impact of LDA dimension on I-vector recognition rate

Table 3 Influence of LDA dimension on i-vector recognition rate

LDA dimension	**EER**
Dim=200	7.03%
Dim=180	6.98%
Dim=150	5.62%
Dim=110	4.82%
Dim=100	4.63%
Dim=90	4.54%
Dim=80	4.78%
Dim=50	5.86%

E. Testing the impact of voice duration

Under the premise of using LMS adaptive filtering denoising algorithm and Mel Frequency Cepstral Coefficients (MFCC) characteristic extraction algorithm, to compare the recognition accuracy of GMM-UBM and i-vector model by using the speech with same durations and different durations. recognition accuracy. Wherein, the number of segments of the read digit string is used to indicate different test speech durations (a string of digits represents 2 seconds).

Table 4 Recognition accuracy of VP master using GMM-UBM and i-vector models for different speech duration

Digits string/segment (%)/ Recognition model	1	2	3	4	5
GMM-UBM	77	80	85	88	90
i-vector	80	86	91	95	97

The experimental data is shown in *Table 4*. From the results, it can be seen that compared with the GMM-UBM model, the used i-vector model can gain higher recognition accuracy with shorter test speech duration. Among them, the recognition accuracy is better when using test input speech of 5-digit string (i.e., about 10 seconds). This is because in the case of using same test speech duration (i.e. the same

training set size), the i-vector model can more effectually cover the dimension of the target speaker's Voiceprint feature than the GMM-UBM model ((that is, it can more represent speaker's Voiceprint feature), so the recognition accuracy of the i-vector model is higher. Moreover, the longer duration of the test speech (i.e., the larger the speech training set), the more coverage of the speaker's Voiceprint feature can be covered, so the accuracy of the model recognition is higher when the test speech duration is longer. How to choose the input duration of the test speech in the product depends on the designer's balance of the inverse relationship between the user experience and the accuracy. In the system, in order to obtain higher accuracy, a 5-digit string is used as the test speech input.

4.Conclusion

The performance of the Voiceprint recognition system is influenced by male to female ratio in the training set and Gaussian mixture model as well as speech length, etc. By comparing experimentation, it is approved that the holistic ERR of recognition system based i-vector is lower than that of a system based GMM-UBM. Furthermore, the needed time is more less while the gained precision is higher. However, the research of this system still needs to further consider the actual user needs, such as how to identify the rate and real-time when the number of training samples is insufficient, which is worth further study.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Zhaohui W, Yingchun Y. Speaker recognition model and method. Beijing: Tsinghua University Press, 2009, pp.14-17.

[2] Rao RR, Prasad A, Rao CK. Robust features for automatic text-independent speaker recognition using Gaussian mixture model. International Journal of Soft Computing and Engineering. 2011; 1(5):330-5.

[3] Drgas S, Virtanen T. Speaker verification using adaptive dictionaries in non-negative spectrogram deconvolution. In international conference on latent variable analysis and signal separation 2015 (pp. 462-9). Springer, Cham.

[4] Swietojanski P, Ghoshal A, Renals S. Convolutional neural networks for distant speech recognition. IEEE Signal Processing Letters. 2014; 21(9):1120-4.

[5] Li Z, HE L, Zhang W, Liu J. Speaker recognition based on discriminant i-vector local distance preserving projection [J]. Journal of Tsinghua University (Science and Technology). 2012.

[6] You CH, Li H, Ma B, Lee KA. A study on GMM-SVM with adaptive relevance factor and its comparison with i-vector and JFA for speaker recognition. In international conference on acoustics, speech and signal processing 2013 (pp. 7683-7). IEEE.

[7] Gupta V, Kenny P, Ouellet P, Stafylakis T. I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In international conference on acoustics, speech and signal processing 2014 (pp. 6334-8). IEEE.

[8] Cumani S, Laface P. Scoring heterogeneous speaker vectors using nonlinear transformations and tied PLDA models. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018; 26(5):995-1009.

[9] Kanagasundaram A, Dean D, Sridharan S, Gonzalez-Dominguez J, Gonzalez-Rodriguez J, Ramos D. Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Communication. 2014; 59:69-82.

[10] Lu X, Shen P, Tsao Y, Kawai H. Regularization of neural network model with distance metric learning for i-vector based spoken language identification. Computer Speech & Language. 2017; 44:48-60.

[11] Wang W, Xu J, Yan Y. Identity vector extraction using shared mixture of PLDA for short-time speaker recognition. Chinese Journal of Electronics. 2019; 28(2):357-63.

[12] Ahmed AI, Chiverton J, Ndzi D, Becerra V. Channel variability synthesis in i-vector speaker recognition. IET international conference on intelligent signal processing 2017.

[13] Nayana PK, Mathew D, Thomas A. Performance comparison of speaker recognition systems using GMM and i-vector methods with PNCC and RASTA PLP features. In international conference on intelligent computing, instrumentation and control technologies 2017 (pp. 438-43). IEEE.

[14] Joy NM, Kothinti SR, Umesh S. FMLLR speaker normalization with i-vector: In pseudo-FMLLR and distillation framework. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018; 26(4):797-805.

[15] Xu L, Lee KA, Li H, Yang Z. Generalizing i-vector estimation for rapid speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018; 26(4):749-59.

[16] Al-Kaltakchi MT, Woo WL, Dlay SS, Chambers JA. Speaker identification evaluation based on the speech biometric and i-vector model using the timit and ntimit databases. In international workshop on biometrics and forensics 2017 (pp. 1-6). IEEE.

[17] Kanagasundaram A, Dean D, Sridharan S, Ghaemmaghami H, Fookes C. A study on the effects of using short utterance length development data in the design of GPLDA speaker verification systems. International Journal of Speech Technology. 2017; 20(2):247-59.

[18] Lizhe T, Dawei F, Dongsheng L, Rongchun L, Feng L. Analysis of large-scale distributed machine learning systems: a case study on LDA. Journal of Computer Applications. 2017; 37(3): 628-34.
[19] Lei L, Kun S. Speaker recognition using wavelet packet entropy, I-Vector, and Cosine Distance Scoring. Journal of Electrical and Computer Engineering. 2017.

Jing Zhang received the B.E. degree from Shenyang Ligong University, Shenyang, China, in 2000, and the M.Sc. and Ph.D. degrees from the Guangdong University of Technology, Guangzhou, China, in 2003 and 2015, respectively. Since 2003, she has been with the Guangdong University of Foreign Studies, Guangzhou. Her current research interests include Digital Signal Processing and Artificial Intelligence.
Email: ha_go@163.com