Factor Analysis Method for Text-independent Speaker Identification

Tingting Liu
University of science and technology of China, Hefei, China
Email:ltt1989@mail.ustc.edu.cn

Shengxiao Guan
University of science and technology of China, Hefei, China
guanxiao@ustc.edu.cn

***Abstract*—Factor analysis method offers state-of-the-art performance in speaker identification during the paper. The compact representations of speakers named i-vectors are extracted from the utterances in a new low dimensional speaker- and channel-dependent space, named a total variability space. LBG algorithm is combined with fuzzy theory in the initialization of speaker models,which improves the recognition rate of the system. Channel compensation techniques, such as Linear Discriminate Analysis (LDA), Principal Component Analysis (PCA), Nuisance Attribute Projection (NAP) and Within-class Covariance Normalization (WCCN) are compared during the experiment. It can be seen that LDA followed by WCCN achieves satisfying performance. In addition, several identification methods are contrasted in the experiments. One is through Support-Vector-Machine (SVM), another one directly uses the cosine distance similarity (CDS) as the final decision score, logarithmic likelihood and vector quantization are used to compare to above two methods. It demonstrates that CDS combined with score normalization obtains better result. The testing of mobile phone database shows the robustness of the system in complex channel environment. The graphical user interface of training and testing module is simulated on MATLAB in the end of the paper.**

***Index Terms*—factor analysis, total space, i-vector, channel compensation, cosine similarity, score normalization**

I. INTRODUCTION

Speaker recognition is becoming an increasingly significant biological authentication technology. Speaker identification mainly aims to identify the exact speaker from speaker utterance corpus. For the speech of high quality, such as the consistent training and testing background environment, general systems can achieve high recognition rate. Yet, the complexity and peculiarity of the transmission channels or other interference factors will make the performance of the system degrade sharply. Thus the systems cannot adapt to the development of practical environment, which leads to our research on the robustness of speaker recognition. Channel compensation or score normalization could be applied in order to isolate the target speakers.

Joint Factor Analysis (JFA), originally proposed by Kenny [1], has received considerable focus due to its successful application to the speaker recognition task. Both the channel effects and the information about speakers are contained in the speech. The first step of factor analysis is to train the Gaussian Mixture Models (GMMs). GMMs were obtained by adapting the Universal Background Model (UBM) with the use of the algorithm of Maximum A Posteriori (MAP) [2][3]. Then Gaussian Mixture Super Vectors (GSVs) are produced by concatenating the means of GMMs, which are used to train the eigenvoice and eigenchannel space. JFA assumes each GSV is composed of two independent components, one of them is associated with the speaker itself, and the other one has relationship with channel.

More recently, i-vector approach motivated by JFA was introduced. It only defines one variability space termed total variability space, instead of joint estimation of separate speaker and session spaces and factors. I-vectors stand for sessions of utterance in the new total space. Since the channel variability information is included in this total variability space, i-vectors need to be in conjunction with channel compensation methods, for example, within-class covariance normalization, linear discriminant analysis, principal component analysis and nuisance attribute projection. In the paper, we contrast these compensation techniques, and find that the best result comes from the combination of LDA and WCCN [4]. This algorithm makes full use of the advantage of LDA and WCCN, which has the maximum disparity as well as the minimum overall cost.

The main content of the paper is showed as follows. In section II, we describe the total variability space and i-vectors. Section Ⅲ describes the channel compensation techniques. We analyze the recognition methods in section IV.The simulation experiments and results analysis are given in section V. Section VI is the conclusion of the paper.

doi:10.4304/jsw.9.11.2851-2860

II. Factor Analysis

This section mainly describes the following stages, including GMM training, total variability space training, i-vectors acquiring [5].

A. Training Gaussian Mixture Models

The first step is to initialize the clustering centers after extracting the feature parameters of all the utterances from different speakers. Then we apply LBG algorithm followed by fuzzy theory to get the new clustering centers. LBG algorithm was first proposed by Linda, Buzo and Gray in 1980. Since the approach does not take the prior information into account, it may lead to a negative influence on the final cluster centers. Thus fuzzy theory is used to update the clustering centers. The following formula is the definition of the function of overall cost.

$$J_m = \sum_{t=1}^{T}\sum_{j=1}^{M}(\mu_{jt})^m d^2(x_t, c_j) \tag{1}$$

$$d^2(x_t, c_j) = \left\| x_t - c_j \right\|^2 = (x_t - c_j)^T F_j^{-1}(x_t - c_j) \tag{2}$$

where M is the order of the GMM. μ_{jt} is the membership, which ranges from zero to one. Initial cluster centers are obtained when the partial derivatives are zero.

$$c_j = \sum_{t=1}^{T}(\mu_{jt})^m x_t \Big/ \sum_{t=1}^{T}(\mu_{jt})^m \tag{3}$$

In order to estimating the models accurately, we should utilize EM algorithm to obtain the optimal centers. Above procedures are used to acquire the UBM. UBM represents the speaker- and session-independent characteristics of all the speakers. Experimental results have shown that this initialization approach improves the accurate recognition rate to some degree.

To train GMMs corresponding to UBM is the next step. GMM is acquired by the adaptation of UBM. Then Gaussian Supervectors (GSVs) are produced by connecting means of GMMs. This self-adaption is only used to update mean vectors of GMMs; however, we assume that all the GMMs have the same weigh vector and covariance matrix.

B. Total Variability Space

In JFA system, we need to estimate both the channel and the speaker space. The speaker space is defined by the eigenvoice matrix V and the channel space is represented by eigenchannel matrix U. In the eigenvoice training, all the recordings of a given speaker are considered to belong to the same person. Meanwhile, in the eigenchannel training we should get the utterances in different channels. The training data is so complex that we begin the research on total space. The process of training total variability space is exactly the same as eigenvoice matrix training. Both the feature of speaker and the channel variability are included in the space.

It has to be mentioned that the entire utterances are produced by different speakers. The space assumes that an utterance can be represented by GMM supervector, which is concatenated by the mean vectors of each GMM. A Gaussian supervector M is defined as

$$M = m + T\omega \tag{4}$$

where m can be replaced by UBM supervector. The matrix T is the definition of the total space. Moreover w is a random vector, whose distribution can be described as $N(0, I)$. w is also called identity vector, which is referred to as i-vector in short. M is normally distributed. The mean vector of M is m and covariance matrix of M is TT^T. The model of i-vector can be seen as the reduction of dimension of the GMM supervector, which projects the GMM supervector onto the total variability space. Since the dimension of total variability space is far smaller than that of supervector space, the process makes the following manipulations such as intersession compensations, scoring, become tractable. The algorithm of training total variability space matrix T is given as follows

Step1: Compute the Baum-Welch statistics for the given utterance h. The statistics are extracted using UBM, similar to [6].

$$N_{c,h}(s) = \sum_t \gamma_t(c) \tag{5}$$

$$\widetilde{F}_{c,h}(s) = \sum_t \gamma_t(c)(Y_t - m_c) \tag{6}$$

$$S_{c,h}(s) = diag\left\{ \sum_t \gamma_t(c) Y_t Y_t^T \right\} \tag{7}$$

where c is the Gaussian index. $\gamma_t(c)$ corresponds to posterior probability of mixture component c generating the vector Y_t. The centralized Baum-Welch first order statistics are showed in equation 3.

Step2: The initial value of matrix T is produced randomly. EM algorithm is used as an iterative method to estimate the total variability space matrix.

Step3: Compute the posterior distribution of the hidden variable $\omega_{s.h}$.

$$l(s) = I + T^T \Sigma^{-1} N_h(s) T \tag{8}$$

$$E\left[\omega_{s.h}\right] = l^{-1}(s) T^T \Sigma^{-1} \widetilde{F}_h(s) \tag{9}$$

$$E\left[\omega_{s.h}\omega_{s.h}^T\right] = E\left[\omega_{s.h}\right] E\left[\omega_{s.h}^T\right] + l^{-1}(s) \tag{10}$$

Where $N_h(s)$ is defined as the diagonal matrix of a given utterance h of speaker s, whose diagonal blocks are $N_{c,h}(s)I$. $\widetilde{F}_h(s)$ is a supervector obtained by concatenating all the $\widetilde{F}_{c,h}(s)$ for a given utterance h. Σ can be replaced by the covariance matrix of UBM.

Step4: Recalculate the statistics below through the data of training set and acquire the maximum likelihood.

$$\Phi_c = \sum_s \sum_h N_{c,h}(s) E\left[\omega_{s.h}\omega_{s.h}^T\right], (c = 1, 2, \cdots C) \tag{11}$$

$$\Omega = \sum_s \sum_h \widetilde{F}_h(s) E\left[\omega_{s.h}^T\right] \tag{12}$$

The updating formula of total variability space matrix T is written as follows.

$$T_i\Phi_c = \Omega_i \ (i = 1,2,\cdots CP) \tag{13}$$

In general, the number of iterations of EM algorithm is about ten. In the paper, we decide to concatenate all the recordings of each speaker into one utterance. That is to say, each speaker has a whole session of speech. The value of the variable h is defined as one. In this case, the procedure of training the total variability space is simplified to some degree.

C. Identity Vector

Identity vector/total factor $\omega_{s.h}$ is a hidden variable, whose posterior distribution is also a Gaussian distribution. It has to mention that mean vector of total factor is the estimation of i-vector. Three steps are needed to obtain i-vectors, which is specifically described below

Step1: Calculate the Baum-Welch statistics of each target speaker.

Step2: Read the trained total variability matrix T.

Step3: Compute the mean value of $\omega_{s.h}$.

The estimation of i-vectors is showed in equation 9. The goal of factor analysis is to realize the extraction of low-dimensional features. Basically, i-vectors are the new features of each speaker. In the actual experiment of the paper, we don't carry out the process of subsection on the training utterances of each speaker. That is to say, consider the training speech of one speaker as one session. The following experiment proved that it doesn't affect the recognition rate of the system, instead, improves the efficiency of acquiring i-vectors.

III. Channel Compensation

As the total variability space, which is represented by matrix T, contains both speaker and channel variability, additional intersession/channel compensation techniques are required. It is an integral part of speaker recognition task, which can significantly reduce the classification errors. Three compensation techniques are described in this section as applied to i-vectors: Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Nuisance Attribute Projection (NAP), and Within Class Covariance Normalization (WCCN).

A. Linear Discriminant Analysis

Due to the reduction of dimension, LDA is extensively used in the scope of pattern recognition [7][8]. LDA helps better discriminate between separate classes through finding out the orthogonal axes of total space. The entire speech of one speaker is considered as one class. LDA algorithm attempts to simultaneously maximize the inter-speaker discrimination and minimize the intra-speaker variance. Therefore, the problem of optimization can be turned into maximizing the Rayleigh coefficient below

$$J(y) = \frac{y^T S_B y}{y^T S_w y} \tag{14}$$

The equation showed above describes the significance of the Rayleigh coefficient. And the between-class variance S_B and within-class variance S_W are calculated respectively as follows

$$S_B = \sum_{s=1}^{S} (\overline{\omega}_s - \overline{\omega})(\overline{\omega}_s - \overline{\omega})^T \tag{15}$$

$$S_W = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{h=1}^{n_s} (\omega_{s.h} - \overline{\omega}_s)(\omega_{s.h} - \overline{\omega}_s)^T \tag{16}$$

The mean vector of total factor is supposed to a zero vector since the identity vector is a random variable with a standard normal distribution. However, we use the actual computed global mean vector rather than assuming it was zero. The goal of maximization is to acquire a projection LDA matrix A which is constituted by the eigenvectors with the highest eigenvalues. The equation is written below

$$S_B y = \lambda S_w y \tag{17}$$

We can choose the k eigenvectors having the best eigenvalues given in equation 14 to construct the LDA matrix. LDA helps project the total factors onto a low-dimensional space.

B. Principal Component Analysis

Principal component analysis [9] is commonly used in data compression. The direction of projection is obtained by maximizing the projection value of feature. Detailed computing process is described as follows

Step1: Calculate the mean of i-vectors of every speaker. Then the intermediate value is obtained by difference between i-vector and the mean.

Step2: Acquire the covariance matrix of these above characteristics which are described in the following formula.

$$\mathrm{cov} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1R_\omega} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2R_\omega} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{R_\omega 1} & \sigma_{R_\omega 1} & \cdots & \sigma_{R_\omega R_\omega} \end{bmatrix} \tag{18}$$

Step3: Compute the eigenvalues and eigenvectors of covariance matrix. Then the next step is to normalize the eigenvectors.

Step4: Sort the eigenvalues in descending direction. Then the first K values corresponding to the eigenvectors are constructing the projection matrix.

C. Within Class Covariance Normalization

Attenuating the dimensions of high within-class variance is the purpose of WCCN [10]. However, it guarantees only the conservation of directions in space, and removes information about the between-class variability. We make the assumption that all the voices belong to one category of a given speaker. The matrix of WCCN is calculated as below

$$W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega_{s.h} - \overline{\omega}_s)(\omega_{s.h} - \overline{\omega}_s)^t \tag{19}$$

where $\overline{\omega}_s$ is the mean value of i-vectors of the speaker s, and S is the number of speakers. n_s is the number of utterances of speaker s. Considering with the form of cosine kernel, a feature projection function can be defined as follows

$$\varphi(\omega) = B^T \omega \tag{20}$$

Where B is obtained through Cholesky decomposition of the inverse of the within class covariance matrix $W^{-1} = BB^t$. WCCN applies the within class covariance matrix to normalize the cosine kernel function aim to realize channel compensation.

D. Nuisance Attribute Projection

NAP needs to construct a feature space that mainly describes the information of channel [11]. The purpose is to seek for the space which can best describe the characteristics of speakers. It allows us to project i-vectors onto speaker space in order to remove the nuisance direction. The projection matrix is showed in the following formula.

$$P = I - RR^T \tag{21}$$

where R is a matrix of low rank. The number of column vectors is k. The matrix is obtained by the eigenvectors of within-class covariance matrix with the first k large eigenvalues showed in equation 19. What is more, the channel space is assumed to be made up of all these eigenvectors.

IV. Recogniton Method

The system based on factor analysis applies several methods to estimate the similarity. Vector quantization is described in the previous research work. Thus we mainly describe support vector machine (SVM), cosine distance scoring (CDS) and logarithmic likelihood. Score normalization is used to reduce the difference caused by channel variability.

A. Support Vector Machine (SVM)

Using i-vectors to train the SVM has several advantages relative to GMM supervector, such as lowering the dimensions of features, providing more convenience for modeling, reducing the amount of computation in channel compensation. In general, SVM is mainly applied in two-class classification. We should design a series of two-class classifiers to solve this multi-class classification problem. For example, if N speakers are included in the speaker corpus, the number of classifiers is C_N^2. These multi-class classifiers are trained by all the utterances of the speakers.

The weakness of this approach is that these classifiers may interfere with each other and then have a negative impact on the recognition rate. Therefore we may attempt to take advantage of global traversal algorithm. That is to say, i-vectors of testing utterance are as the input to every classifier, and then calculate the score acquired by each classifier. As a result, the class having the highest score is the target speaker.

In this paper, we use a library of support vector machine (LIBSVM) [12], which is currently one of the most widely used SVM software. Two steps are included in the typical use of LIBSVM: firstly, to obtain a model by the training dataset and secondly, to predict information of testing data by the known model. The tool package provides us with different data formats and SVM parameters to choose.

B. Cosine Distance Scoring

Comparison between channel compensated i-vectors for speaker identification can also be accomplished using other approach. As both training and testing i-vectors undergo the same transformation，cosine distance [13] can be seen as a symmetric classification method. Given two i-vectors, ω_{target} from a known speaker, and ω_{test} from a unknown speaker, the cosine similarity is calculated as follows

$$score(\omega_{target}, \omega_{test}) = \frac{\langle \omega_{target}, \omega_{test} \rangle}{\|\omega_{target}\| \|\omega_{test}\|} \tag{22}$$

Note that the cosine distance scoring (CDS) only considers the angle between the two i-vectors and not their magnitudes. It is believed that channel/session information is included in the magnitude. Thus cosine distance scoring is reasonable to identify the target speaker in the utterance corpus and can improve the robustness of the system to some degree.

C. Logarithmic Likelihood

Logarithmic likelihood refers to the calculation of the probability of each speech frame. The object speaker is the one who gets the highest score. The probability value is computed below

$$P(\lambda_k | X) = P(\lambda_k | x_1, x_2, \cdots, x_T) \tag{23}$$

According to the Bayesian formula, the above formula can be converted to calculating $P(X|\lambda_k)$. If the probability is so small, the numerical value may beyond the scope of MATLAB. Therefore the logarithmic value is applied to remove the multiplicative noise and normalize the range of the value.

$$\ln(P(X|\lambda_k)) = \ln(\prod_{t=1}^{T} P(x_t|\lambda_k)) = \sum_{t=1}^{T} \ln(P(x_t|\lambda_k)) \tag{24}$$

As the length of testing speech is not consistent, the logarithmic likelihood should be normalized in order to reduce the impact of voice length on recognition rate. That is to say, the average logarithmic likelihood should be computed in the following formula.

$$score_k = \frac{1}{T} \sum_{t=1}^{T} \ln(\sum_{j=1}^{M} \omega_j P(x_t | j, \lambda_k)) \tag{25}$$

According to the similarity between the testing speech and each speaker in the database, the one corresponding to the maximum likelihood is the target speaker.

D. Score Normalization

Score normalization aims to counteract statistical variation in classification scores. This is accomplished by

scaling all the scores to a global distribution with a zero mean and unit variance. Recently, score normalization technique is becoming inevitable in speaker recognition system [14].

In this paper, we decide to use Z-norm to normalize the scores. Z-norm is a kind of normalization technique; it aims to estimate the mean and variance of the score, and perform linear transformation to the score.

Compared to Z-norm, there are T-norm, and ZT-norm. Combination of different normalization techniques may not have a superimposed effect. Considering that T-norm needs to compute the mean value and covariance of several testing speech to normalize the score, thus the procedure will cost more time which may have a negative influence on the real-time requirement of the system. Therefore, the simple and convenient Z-norm is applied in the final score normalization process. The specific process of Z-norm is showed in the figure above.

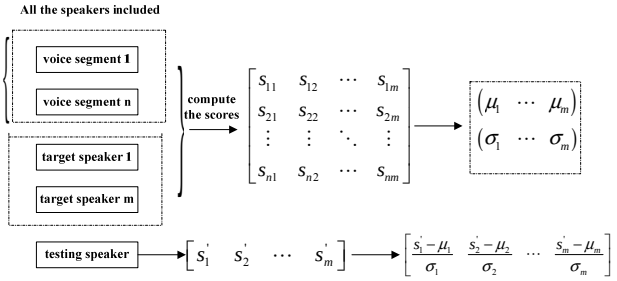

Figure 1. The flow chart of score normalization.

V. Experiments And Results

A. Corpora

The experiments are performed on the corpus designed by the members of our laboratory to evaluate the text-independent speaker identification system. The corpus includes the recordings of 50 speakers in all (30 males and 20 females). These voices are recorded at an 8 KHz sampling rate with 16bit-quantization precision in the general laboratory environment on the same microphone. To each speaker, the length of the training data lasts 3 minutes, and the length of testing data is 30 seconds. Then we divide the training data into 40 sessions, and the testing data into 10 sessions. To make sure that each speaker has 40 samples to train the identification model and 10 samples to identify the target speaker. That is to say, we have 500 samples to test the accuracy of the system. It is mentioned that generally each sample will be divided into several sessions, each of which lasts about two seconds.

The experiments were carried out to compare the influence of different compensation techniques on recognition rate. On the other hand, we apply separate classification methods, such as SVM, cosine distance scoring, vector quantization and logarithmic likelihood. In the experiments, we will contrast the result of SVM with that of cosine distance. Despite of this point, we use the same channel compensation for these two conditions.

Another corpus used in our experiments is from a voice library of MIT mobile phone speaker recognition [15]. It is mainly conducted to test the performance of the system in different recording devices in order to show the advantage of i-vector based speaker identification system.

B. Experimental Configuration

Cepstral parameters, which are acquired by hamming window with the length of 30ms, are used in the experiments. The shifting is 15 ms. 12 Mel Frequency Cepstral Coefficients (MFCC) with appended delta coefficients were calculated. Two gender-dependent UBMs with 32 Gaussian Mixture Models were trained on microphone and mobile phone data using Expectation Maximum (EM) algorithm. The Random method is used to initialize the cluster centers. The training voices were truncated into several interval periods in order to increase the amount of the development data. Then GMM of each speaker was acquired by combining UBM with MAP. Concatenating the mean vectors of GMM is to acquire the GMM mean supervector, which is used to train the total variability matrix.

The dimension of i-vectors is generally 400. The corpora are also applied in training LDA, PCA, WCCN, NAP matrix to accomplish the channel compensation. Z-norm was utilized for the procedure of score normalization. LDA is the projection to reduce the dimension; however, WCCN and NAP are the equal dimension mapping. We assume the dimension of i-vector is 300 after the procedure of LDA. SVM training and classification was performed using i-vectors after session compensation. Cosine distance is used to compare the performance with SVM. Where applicable, score normalization is employed by Z-norm.

Evaluations in this paper focus primarily on three aspects: Firstly, the paper discusses the comparison between different compensation techniques as well as the impact of LDA and PCA dimension selection on recognition rate under the system of different orders. Then it comes to the influence of initialization method on accurate recognition result. Secondly, it contrasts separate identification approaches in speaker recognition system. In the end, the paper concludes the robustness of the speaker identification system based on factor analysis approach and its application on the utterances of mobile phone. It has to mention that all the experiments are realized on the software of MATLAB R2009b.

C. Results

All our experiments were carried out on the corpus of 50 speakers, 35 males and 15 females are included. Another corpus is the MIT mobile phone utterance library, normally 3 min in training and 30 seconds in testing.

1) The experiments carried in this section compare the recognition rates under several conditions. The goal of the first experiment is to contrast the performance of LDA with that of PCA in removing the nuisance directions. The results are reported in Figure 2. It's worth noting that during the stage of matching we make use of cosine distance scoring without score normalization.

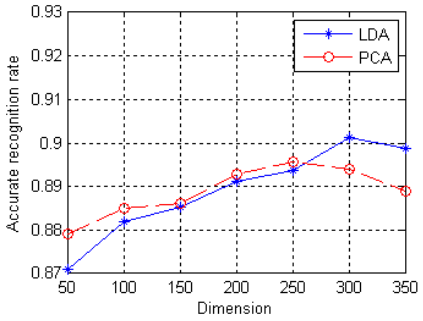

Figure 2. The effect on accuracy by .LDA and PCA in different dimension.

The results show that when the dimension is low, the effect of PCA is better than that of LDA. However as the increase of dimension, especially higher that 250, the system based on LDA behaves better. Considering both the recognition rate and experimental result, LDA is chosen and the optimal dimension of LDA space is 300 from the figure. LDA helps rotate the space in order to minimize the intra-speaker variance; and improve the performance in the case of cosine kernel. During the training step, we train the LDA projection matrix on all utterances used for training T matrix. The training data is projected onto the low-dimensional space to compensate for channel effects.

After determining the influence of dimension, it comes to the research on different order GMM in LDA based systems. The comparison result is showed in the below figure.

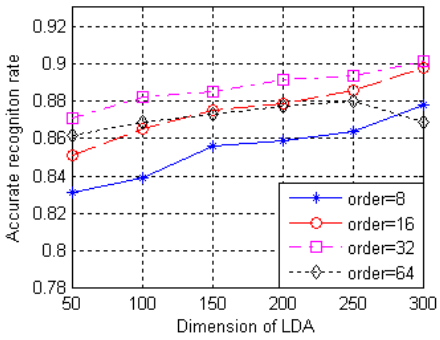

Figure 3. The effect on accuracy by LDA under the system of different order..

It can be seen in the figure that the recognition rate is on the rise as the order of the system becomes higher. The trend is embodied in the order of 8, 16 and 32. When the order rises to 64, there will be a certain degree of decline. The increase of the order will make the interval between different curves become reduced. That is to say, the model cannot well represent the characteristics when the dimension is too high or too low. If the dimension is too low, the models cannot distinguish among the speaker. However, the space dimension disaster will occur if the dimension is too high.

From the experiments, we can find the marked improvement with channel compensation. Moreover, single compensation technique will not appropriate for intersession compensation. Thus the comparison of accuracy between separate compensation techniques is showed in the following Table Ⅰ.

The experiments were carried on the same database. According to the results showed in the above table and figure， we note that applying WCCN in the LDA-projected space helps to improve the performance as compared to other single channel compensation techniques such as PCA, WCCN, NAP. Moreover, it is obvious that WCCN depicts better results than NAP, which makes us think over the combination of NAP and WCCN. Although, to some degree, the integration of

TABLE.I
COMPARISON BETWEEN DIFFERENT COMPENSATION TECHNIQUES

Compensation technique	Accuracy rate
LDA(dim=300)	90.1%
PCA(dim=300)	89.2%
WCCN	89.4%
NAP	88.9%
NAP+WCCN	90.8%
LDA(300)+WCCN	91.6%
LDA+NAP+WCCN	89.3%

WCCN and NAP improves the ratio, it does not obtain the best performance, so does the joint of three channel compensation techniques. Above all, the combination of LDA with WCCN achieves the best accurate recognition rate. The score of cosine distance after LDA and WCCN is showed in the following formula.

$$\hat{\omega}_{LDA\_WCCN} = B^T A^T \omega \tag{26}$$

TABLE.II
COMPARISON OF DIFFERENT INITIALIZATION APPROACHES

Initialization	random	LBG	LBG+Fuzzy
Accuracy	91.6%	92.1%	92.7%

The purpose of WCCN approach is to compensate the intra-speaker variability. Theoretically speaking, the application of WCCN in the two-dimensional LDA space reduces channel effects by minimizing the within-speaker variance. It is significant to note that the matrix of WCCN is a diagonal matrix after two dimensional LDA and WCCN projection. Thus the distribution of samples only executes the scale transformation without the change of rotation.

From the above experiments, it can be seen that LDA followed by WCCN acquires satisfying result. Then next experiment will show the contribution of initialization method on recognition rate, which increases by 1.1%. The following table II shows the result of different initialization methods including random, LBG algorithm, and LBG algorithm followed by fuzzy theory.

2) In this section, the thesis discusses difference between different identification approaches, including support vector machine, cosine distance scoring, logarithmic likelihood and vector quantization. It has to

mention that LBG algorithm combined with fuzzy theory is applied to generate the initial cluster centers.

Previously, SVM is always combined with JFA to constitute the system of JFA-SVM. The method of SVM is mainly through training SVM classifiers and then we have to calculate the score of each speaker. The target is the speaker with the highest score. We use the software package of LIBSVM as a library for support vector machines in the paper. In order to use this package, we have to change the data set into spectacular form. The form of training dataset and testing dataset is as follows

<label> <index> : <value1> <index>: <value2>…

Label is to identify the categories of the utterances. Index represents the numerical order of the exacted features. The index number starts from one. Values stand for every dimensional value of i-vectors after intersession/channel compensations. Then we should refer to three principal functions. One of them is libsvmread, which helps reading the testing and training data. The function of svmtrain is to produce a model for solving an optimization problem of SVM. The last one is svmpredict, which utilizes the trained model to get the category of the testing data. The advantage of the package is that we can choose different kernel functions and other parameters. [16].

The approach of cosine distance scoring is also applied in the total variability space. Compared to SVM, cosine distance scoring (CDS) provides a similar performance with a considerable increase in efficiency. The cosine similarity score operates by comparing the angles between the testing i-vector and the target i-vector after channel compensation without considering the amplitudes of i-vectors. The experiments are performed on the same dataset. The scores are normalized with Z-norm which further compensated for nuisance effects. The parameter to evaluate the system is the accurate recognition rate [17][18].

The method of logarithmic likelihood is to compute the probability scoring of testing speech sequences in each GMM. The target is the model with the best score. Vector quantization is to think i-vectors after channel compensation techniques as the parameters, which are used to generate the best codebook. The size of the codebook is 64. After quantizing the feature vectors of testing sequence, the relative difference between the testing speech and codebooks should be calculated. The speaker with minimum error is the identification result. The performance of the identification system using the above four method is displayed in table Ⅲ.

TABLE.III
COMPARISON BETWEEN DIFFERENT IDENTIFICATION METHODS

Method	I-vectors	LDA	LDA+WCCN
VQ	79.4%	82.5%	84.2%
SVM	87.8%	86.5%	89.8%
LLR	87.5%	88.6%	91.2%
CDS	89.1%	91.6%	92.7%
CDS+Znorm	89.8%	92.3%	93.4%

From the above table, we can see that the system that combines LDA with WCCN using cosine distance scoring achieves the highest accurate recognition rate. Meanwhile, it is obvious that within class covariance normalization definitely optimizes the classification performance of SVM [19]. In a word, cosine distance scoring improves not only the accuracy of the system, but also the efficiency of algorithm to some degree [20].

TABLE.IV
AVERAGE TIME OVERHEAD OF DIFFERENT METHODS

Method	VQ	SVM	LLR	CDS
time	4.1s	6.7s	10.2s	8.6s

The above table IV shows the average time overhead of these identification methods. Vector quantization [21][22] method is the fastest approach relative to others. The second is support vector machine; logarithmic likelihood costs the highest time consumption. Because it needs to calculate the probability value of each frame of testing speech relative to different speaker, which increases the complexity of the process. It can be seen that cosine distance scoring is the modest weighing the algorithm complexity and identification rate.

3) This section illustrates that factor analysis method improves the robustness of the system in speaker identification. In practical application, a number of external factors have a negative effect on the recognition rate. For example, the training circumstance doesn't match the testing environment; the source of utterances comes from several ways. The experiments carried out in this section are used to verify that the system based on factor analysis can guarantee the recognition rate in spite of using separate sources. LBG algorithm combines with fuzzy theory in the process of initializing models; meanwhile cosine distance similarity with scoring normalization is used in obtaining the final score.

Channel compensation techniques exactly improve the accurate recognition rate of the system as expected in Figure 4. Firstly, as discussed in the paper, the accuracy achieves the highest rate when the dimension of LDA is 300. According to the process of getting i-vectors and the following channel compensations, the dimension of LDA is equal to that of recognition model. Secondly, for the sources come from microphone, the change of recognition rate is not very obvious under the circumstance of LDA and WCCN, which is increased by 3.6%. Nevertheless, for the sources come from microphone, the recognition ratio increased by 4.2% after channel compensations. Thus we can see that i-vectors compensated by LDA and WCCN can represent the characteristics of speaker preferably.

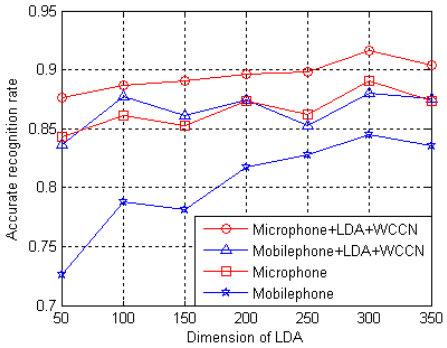

Figure 4. Comparison of accuracy between different sources.

4) The whole process of identification in the paper is simulated on MALAB. The system consists of training and testing module. The GUI of the system is showed in the below Figure 5 and Figure 6.

The first part of the system is training module showed in Figure 5. The speech of every speaker is analyzed at a 30 ms frame length and a 15 ms overlap. The feature parameters are included by 12 MFCCs together with first-order differential coefficients [23]. The GMM of each speaker is trained by the total features of one person. LBG algorithm and fuzzy theory are applied to initializing clusters, and EM algorithm is to update the GMM. GMM supervectos (GSV) are linked by the mean vectors of each GMM component. I-vectors are acquired by projecting the GSVs onto the total variability space. Then we move on to channel compensations. The process of training has ended so far.

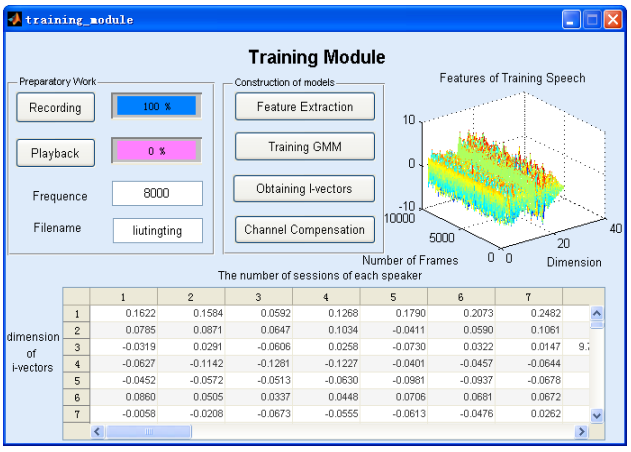

Figure 5. Training Module.

In the training module, the training time of each speaker lasts for 3 minutes, which is divided into 40 sections. We can acquire 40 samples of every speaker as training data in this way. The button of playback it to test that if the microphone is working normally. In addition, the number of components of GMM is selected as 32 based on the above experiments. In the below part of the figure, we can see the table that contains i-vectors of trained speaker. The column of the table represents for the number of i-vectors, while the row of it stands for the dimension of i-vectors. Channel compensation includes the optimized combination techniques discussed in the above experiments.

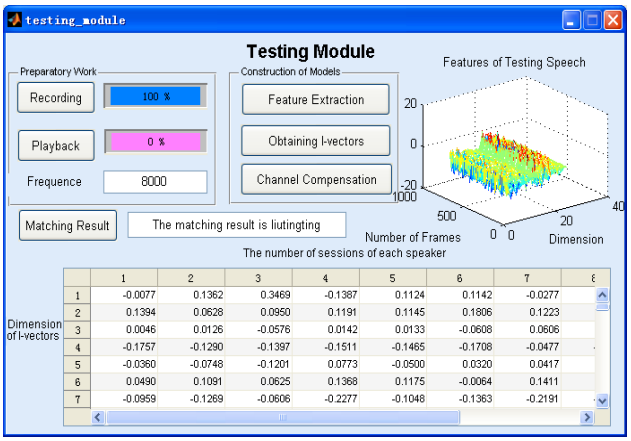

Figure 6. Testing Module.

Another part is testing module showed in Figure 6. I-vectors of testing utterances are obtained during this section according to the trained total variability space, GSVs of testing data, and projection matrix of channel compensations. For example, if the button of channel compensation is pushed down, the projection matrix will be loaded. Then it compares the processed i-vectors of testing data to that of training data using cosine distance scoring (CDS). The speaker with the highest score after score normalization [24][25] is the target speaker.

It is significant to mention that the speech of tested speaker is recorded for 30 seconds, which is changed into 10 partitions. The sum score of these samples will help us find the target speaker displayed in the edit blank showed in the above figure. The right part of the figure displays the feature parameters of the tested speaker. Contrast the feature with that of the target speaker in Figure 5, we can find out that a certain similarity has existed.

VI. Conclusions

A text-independent speaker identification system based on factor analysis is represented in the paper. The method is to define a total variability space, which contains both the speaker and complex channel information. LBG algorithm combined with fuzzy theory is used to initialize the mean vectors of the models. All the GMM supervectors are projected onto this space to obtain the evaluation value of i-vectors. Compensation techniques are applied on i-vectors to remove the channel inference in order to improve the robustness of the system. The paper contrasts several channel compensation techniques, which are respectively nuisance attribute projection (NAP), linear discriminant analysis (LDA), and principal component analysis (PCA), within-class covariance normalization (WCCN). During the experiments in the paper, we can find out that the combination of LDA and WCCN may achieve the satisfying performance. As to the design of the classifiers, in contrast to SVM, LLR and VQ, cosine distance scoring not only provides higher recognition rate, but also makes the decision process less complicated. However, score normalization is inevitable, the goal of which is to counteract statistical variation in classification scores.

From the experiments, the results demonstrated that the system obtains the best performance when the dimension of LDA is 300. The initialization approach mentioned in the paper exactly increases the recognition rate by 1.1% compared to randomly generating the clustering centers or only using LBG algorithm. The technique of LDA may remove the nuisance directions, maximization of the variance between the speakers and minimization of the variance within the speakers. Experiment uses correct recognition rate as the assessment of the systems. Comparing to other approaches, LDA followed by WCCN has shown over 3.6% improvement during the experiment. In addition, the system guarantees the recognition rate in spite of using cellphone utterances, which increase the recognition rate by 4.2%. In the future we intend to optimize the algorithms in order to increase the recognition rate of telephone channel voice. Then the proposed method can be applied in an extended field.

REFERENCES

[1] P. Kenny, P. Quellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," IEEE transactions on Audio, Speech and Language Processing, vol.16, no.5, pp.980-988, 2008.

[2] D. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using Adapted Gaussian Mixture Models," Digital Signal Processing, vol.15, no.4, pp.19-41, 2000.

[3] Gengzhao, Xufei Li, "Identity authentication scheme expansion based on speaker verification," in Journal of computers, vol.8, no.8, pp:2027-2033, 2013.

[4] A. Hatch, S. Kajarekar, and A. Stolcke, "Within Class Covariance Normalization for SVM-Based Speaker Recognition," in international Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 2006.

[5] N. Dehak, R. Dehak, P. Kenny, P. Dumouchel, and P. Quellet. "Front-end factor analysis for speaker verification," In print IEEE trans. Audio, Speech and Language Processing, 2010.

[6] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," IEEE Trans. Speech Audio Processing, vol.13, no.3, May 2005.

[7] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in Proc. Odyssey Speaker and Language Recognition Workshop, 2010, pp28-33.

[8] Bo Yang, Yingyong Bu, "A comparative study on vector-based and matrix-based linear discriminant analysis," in Journal of computers, vol.6, no.4, pp:818-824, 2011.

[9] Ruoyu Yan, Ran Liu, "Principal component analysis based network traffic classification," in Journal of computers, vol.9, no.5, pp:1234-1240, 2014.

[10] D. Matrouf, N. Scheffer, B. Fauve, and J. F. Bonastre, "A straight-forward and efficient implementation of the factor analysis model for speaker verification," in International Conference on Speech Communication and Technology, 2007.

[11] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM supervector Kernel and NAP Variability Compensation," in IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, vol.1, pp.97-100, 2006.

[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] Najim Dehak, Rda Dehak, Partick Kenny, Niko Brummer, Pierre Quellet, and Pierre Dumouchel, "Support Vector Machine versus Fast Scoring in the low-dimensional Total Variability Space for Speaker Verification," in INTER-SPEECH, Brighton, UK, September, 2009.

[14] R. Aukenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for text-independent speaker verification systems," Digital Signal Processing, vol.10, no.1, pp.42-54, 2000.

[15] http://www.datatang.com/data/4143

[16] P. M. Bousquet, D. Matrouf, and J. F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in Interspeech, 2011.

[17] A.Kanagasundaram, R. Vogt,, D.Den, S. Sridharan, and M. Mason, "i-vectors Based speaker recognition on Short Utterances," in Interspeech, 2011.

[18] Mitchell Mclaren, David van Leeuwen, "Source-Normalized LDA for Robust Speaker Recognition Using i-vectors From Multiple Speech Sources," IEEE trans.Audio, Speech and Language Processing, 2012.

[19] J.Weston, C.Watkins, "Multi-class support vector machines," In proceedings of Seventh European Symposium On Artificial Neutral Network, 1999.

[20] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling," Application to speaker verification, Ph.D. thesis.

[21] D. Burton, "Text-dependent speaker verification using vector quantization source coding," IEEE Transactions on Acoustics, Speech and Signal Processing, vol 35, no.2, pp: 133-143, 1987.

[22] F. K. Soong, A. E. Rosenberg, et al, "A vector quantization approach to speaker recognition," AT&T Technical Journal,vol 66, pp:14-26, 1987.

[23] N. Farvardin, R. Laroia, "Efficient encoding of speech LSP parameters using the discrete cosine transformation," International Conference on Acoustics, Speech, and Signal Processing, 1989, pp:168-171.

[24] O.Glembek, L.Burget, N.Brummer, et al, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in IEEE Conference on Spoken Languish, Pittsburgh, PA, USA, September 2006.

[25] N.Dehak, R.Dehak, J.Glass, D.Reynolds, et al," Cosine similarity scoring without score normalization techniques," in Proc. Odyssey Speaker and Language Recognition Workshop, 2010.

Tingting Liu graduate student studies in laboratory of audio-visual Information Processing and Pattern recognition Dep. of Automation, university of science and technology of China (USTC). She received her bachelor's degree in Northeast Forestry University (NEFU) in China. Her research interest is speaker identification. Previously, she mainly focuses on the method of vector quantization. The current study is about speaker recognition system based on i-vectors.

Shengxiao Guan associate professor received PhD degree in engineering from department of Automation in University of Science and Technology of China (USTC). Currently, his research areas concentrate on pattern recognition and intelligence systems, new energy resources and industrial design. His laboratory of audio-visual Information Processing and Pattern Recognition mainly research on intelligent robot,

embedded systems, the control of new energy, speaker identification, image tracking and processing.