PROCEEDINGS OF SCIENCE

Vectoring Gauss Mixture Model Mean Parameters in Speaker Verification

Bo Xu[1]
School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Jiangsu Province
Nanjing, 210023, China
Email: 1094720330@qq.com

Aiyue Chen
School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Jiangsu Province
Nanjing, 210023,China
Email: 1226848849@qq.com

Zijian Shen
School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Jiangsu Province
Nanjing, 210023,China
Email: 616889401@qq.com

In order to realize the text-independent speaker recognition and improve its accuracy rate, a variable method in combination with Gaussian mixture model （GMM） and support vector machine（SVM）is used in this paper. By vectoring GMM mean parameters, we use SVM to recognize the real speaker of the testing speech. We put forward two methods of building hyper plane between two kinds of training speeches and another way to build the hyper plane between Universal Background Model （UBM） and the training speech. Upon vectoring the GMM mean parameters, both of the two methods show better performance in speaker verification than the traditional GMM-UBM withless accuracy. The vectoring GMM mean parameters can amplify the characters of speakers and thus make verification between two speakers in a more obvious manner.

CENet2017
22-23 July, 2017
Shanghai, China

1This study is supported by Jiangsu Academician Innovation and Entrepreneurship（national level）：SZDG2016003

PoS(CENet2017)010

1.Introduction

We take the problem of text-independent speaker verification into account to judge whether the assertion is true or false when a testing speech, the assertion of identity and the corresponding speaker model are given. Simulating the speaker by using an adapted GMM [1] is the standard method.

The use of potential factor analysis to make up for the speaker and channel variability is a stirring part of recent work in GMM speaker recognition. These approaches by simulating the MAP [2] adapted means of a GMM by using potential factors to represent the variation.

We extract the mean parameters of GMM and transit them into high dimension mean vectors in order to build a better environment for SVM [3] and determine the speaker instead of using the Posterior probability calculation directly. Two kinds of training vectors are used in building the hyper plane in which the testing data will be taken. This method is called GMM-SVM. We select Linear Kernel Function[4] to realize SVM classification in the concrete realization process..

For another way if we directly make comparisons between the test vectors and the training data, the calculated amount will increase and the test data is instable to keep close to the true training data. In this way, we choose UBM as a reference because UBM consists of thousands of speech information and it is insensitive to data fluctuation. By using UBM vector and training vectors to build hyper plane, then adapt testing data into calculation and we will estimate who is the speaker of the test data. We call this GMM-SVM-UBM.

During the experiments on the TIMIT database, we use the equal error rate curve to prove that the two modified methods presented in this paper have better recognition ability.

The outline of the letter is as follows. In Section 2, we describe how to vector GMM mean parameters. In Section 3 and 4, we describe two methods for hyper plane building. Finally, in Section 5 and 6, we demonstrate our experiments in TIMIT and give the conclusion of our research.

2.Traditional GMM-UBM and Parameter Extraction of GMM

2.1Pre-processing

For the speech signals, our pre-processing mainly includes two parts. For one thing, calculate Mel Frequency Cepstrum Coefficient (MFCC) [5]. For another, conduct HTK [6] encapsulation. MFCC is to simulate the human auditory system to describe the characteristics of speech signals. The main steps include pre-emphasis, separating frames and window dealing, FFT transform of each frame and filtering, etc.

After the parameter extraction of MFCC and HTK format, the N-dimensions MFCC parameter comprises of N/3 MFCC coefficient, N/3 first order difference parameter, N/3 second order difference parameter and frame energy. It generally has 13 dimensions, 19 dimensions and 40 dimensions, etc.

2.2GMM-UBM

GMM is a multi-dimension probability density function. A GMM consisting of M mixture components can be expressed as weight sum of M gauss members, that is:

$$P(x_i|\lambda)=\sum_{i=1}^{M}\omega_i P(x_i|\mu_i,\epsilon_i) \qquad (2.1)$$

In the above equation, x_i is a D-dimension feature vector; ω_i is the mixture weight of the gauss component.

GMM can be described with each mean vector, covariance matrix and weight of mixture component, thus we can get a GMM parameter. GMM parameter can be calculated through EM algorithm. The discrimination of GMM will be relatively ideal if the time durations of training and testing speeches are both long enough and the speeches are relatively pure.

GMM can't well characterize the speakers when the training speech lasts for tens of seconds while the testing speech only lasts for a few seconds. GMM-UBM firstly obtains a UBM through training speeches of all speakers, then gets the GMM of the target speaker through MAP self-adaption based on UBM. Simulating the uncovered part by using UBM can overcome the shortcomings of GMM well.

UBM is also a large-scale GMM. It is obtained by trainings of a large amount of speech data of different speakers under various environment. UBM is the common reflection of speech characteristics of all speakers and the environment channels. UBM is obtained by training of EM algorithm.

Reynolds [1] proposed the speaker model self-adaption. This method obtains the target speaker model by using the training speech of the target speaker according to MAP. Steps of MAP are as follows:

Given training data of speakers, firstly calculate matching likelihood of each gauss component inand UBM and frame number of the gauss component:

$$P(q_\xi=m|o_\xi,\lambda)=c_m P(o_\xi|m,\lambda)/\sum_{i=1}^{M} c_m P(o_\xi|i,\lambda) \quad (2.2)$$

$$n_m=\sum_{\xi=1}^{\gamma} P(q_\xi=m|o_\xi,\lambda) \quad (2.3)$$

In the above equation, is the frame number of the training speech;is the frame number of the gauss component.

It is proved in numbers of speaker verification experiments that speaker recognition performance is the best if only the mean parameter is modified. So the mean parameter of modified model is calculated according to EM revaluation formula.

$$E_m(S)=\frac{1}{n_{m\,(\xi=1)}}\sum_{\xi=1}^{\gamma} P(q_\xi=m|o_\xi,\lambda)o_\xi \quad (2.4)$$

$$\hat{u_m}=\alpha_m^m E_m(S)+(1-\alpha_m^m)u_m \quad (2.5)$$

In the above equation, $\hat{u_m}$ is the mean vector after modifying; α_m^m is the modifying factor of mean of the gauss component.

2.3Mean Vector Extraction

In the recognition system, if the number of speakers is A, we need to evaluate M groups GMM parameters which represent every speaker's speech information.We select $\hat{\mu}_i$ as a

parameter describes training data. Each $\hat{\mu}_i$ represents an $M \times N$ array. The definitions of M and N are given above. In order to facilitate the use of SVM, we use a vector to express $\hat{\mu}_i$. The value of this vector is $(M \times N) \times 1$, remembering to $\vec{\mu}_i$。Similarly, the mean vector of the testing speech is derived in such a way as. Specific extraction steps can be shown in Figure 1:

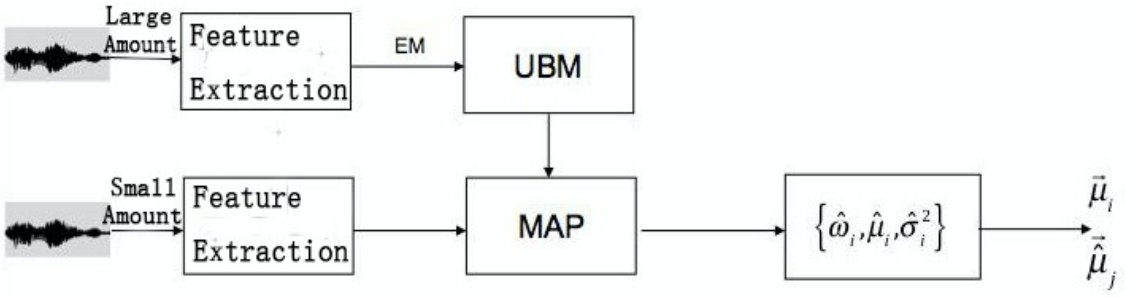

Figure 1: Extraction Process of Mean Vector

3.GMM-SVM

Generally speaking, the sample set (x_i, y_i) can be discussed in the linear separable case. Setting the linear discriminate function as $g(x) = \omega x + b$, in which x_i is a data set, indicating the vector of speaker $\vec{\mu}_i$, y_i is category tag. In order to facilitate the unified labeling, the target speaker in the speaker training model is marked as 1, and the other speaker is labeled as -1. It can be found that the optimal classification function expression can be obtained under the condition of linear separable.

$$f_i(x) = sgn \sum_{i=1}^{k} \alpha_i y_i K(x_i, x) + b \tag{3.1}$$

Among them, a^* and b^* are the optimal classification hyper plane parameters, $(x_i \cdot x)$ is the two vector dot product.

However, as the speech recognition is generally nonlinear separable, it is necessary to map the data to higher dimensions by kernel function. In this paper, we select $K(x, x_j) = (X \cdot X_i)$ the corresponding changes in the inner product space and transform the problem of low dimensional linear non separable two classifications into a linear separable problem.

The result of kernel function is

$$K(X^{(t_1)}, X^{(t_2)}) = \sum_{i=1}^{m} (\sqrt{\lambda_i} \epsilon_i^{-1/2} \vec{\mu}_i^{(t_1)})^T (\sqrt{\lambda_i} \epsilon_i^{-1/2} \vec{\mu}_i^{(t_2)}) \tag{3.2}$$

Where $\vec{\mu}_i^{(t_1)}$ and $\vec{\mu}_i^{(t_2)}$ represent the mean vectors of two kinds of testing speeches, and they build the hyper plane of Function (6).

As the emphasis of this paper is applying GMM parameter to SVM, we did not study the

kernel function selection thoroughly. Expression (6) is the discriminant of classification. $f_i(x)$ corresponds to every speaker's training model. Drag-in testing speech, the calculation can be defined as the degree of deviation of the j test to the i speaker.

The classification progress is represented in Figure 2.

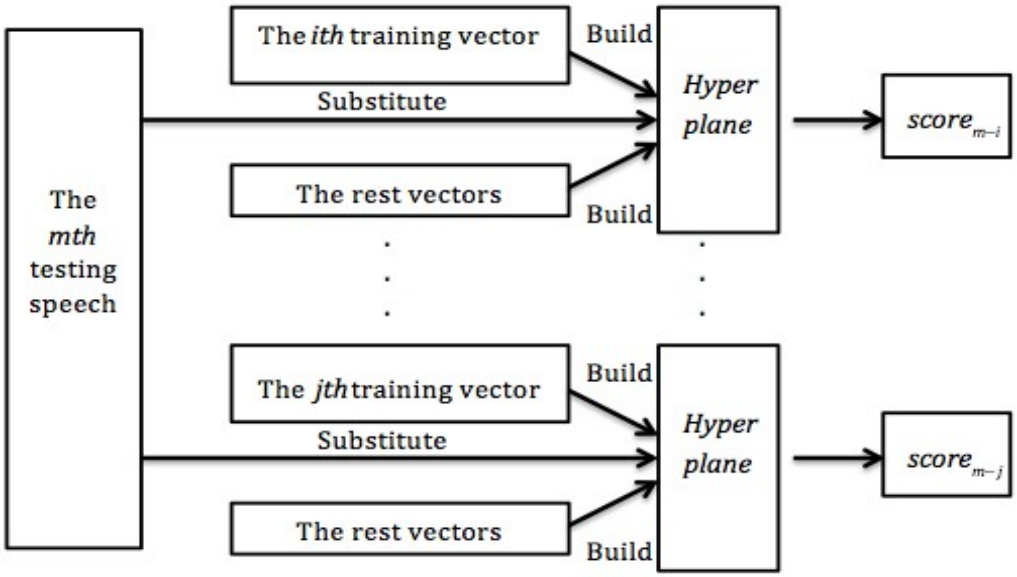

Figure 2: Schematic Diagram(in two dimension) of SVM

4.UBM Reference

Different from GMM-SVM, we also extract UBM mean vector and take it as the testing speech and the reference of the testing speech. Because of such considerations, repeated changes of hyper plane building are needed in calculating the score based on GMM-SVM. As it is time-consuming and likely to be affected by the unbalanced classification, we expect to have a more intuitive solution.

Because of large basic number of the UBM mean vectors, which are almostly not affected by the fluctuation of the data, we set the UBM mean vector as $\vec{\mu}_u$, and the kernel function of (6) will be changed as:

$$K\left(X^{(t_1)}, X^{(t_2)}\right)=\sum_{i=1}^{m}\left(\sqrt{\lambda_i}\epsilon_i^{-1/2}\vec{\mu_u^{(t_1)}}\right)^T\left(\sqrt{\lambda_i}\epsilon_i^{-1/2}\vec{\mu_i^{(t_2)}}\right) \tag{4.1}$$

The hyper plane consists of the character of the UBM mean vector, only one kind of training vector joins in the classification.

In the score calculation based on UBM, the testing speech and UBM will do SVM training together, and get a new classification hyper plane, then take the testing speech into hyper plane calculation to get the score based on UBM. For the principle, the testing speech achieved in the experience achieves quantization through the MAP process with UBM. Therefore, it is close to UBM mean vector in value. When the training speech and UBM get hyper plane through SVM training, the testing speech will lean to UBM so that all scores will have the same dimension and sign consistency and can reflect the difference of value.

5.Experiment Result

GMM-SVM and GMM-SVM-UB, the experimental data in this paper all come from TIMIT voice library with accurate phoneme tagging, hundreds of speakers and thousands of speeches and they were frequently used in speech recognition field. The dimension N of MFCC was 40, in which the energy frame was 1 frame and the speech data of each calculation link were processed according to the 40 dimension.

This paragraph will illustrate the parameter selection in our experiment.

First of all, we choose M in 256, 512 and 1024 respectively.The experiment of GMM-UBM show that the accuracy of verification have its peak value. When we choose M=1024， the dimension of mean vectoring will be $40*1024*1$ The results of GMM-UBM with different M are shown as follows.

M	256	512	1024
Accuracy	82.23%	83.72%	83.95%

Table 1: GMM-UBM by Using Different M

When we use M=1024 as the mixed number in GMM, the other parameters such as UBM, training model and testing model are all using the GMM by 1024 of mixture. I will illustrate the data when we use them in verification and explain the structure of TIIMT in Table 2.

UBM	Training data	Testing data	Scoring criterion
Number of speech: 5300（including 3680 male speech and 1620 female speech） Duration of each speech： 2s	Number of speech： 1000 （ including 70 male participants and 30 female participants ， each participant has ten speeches to train ） Duration of each speech： 2s	Number of speech： 100 （ including 70 male speech and 30 female speech） Duration of each speech： 2s	Each testing speech compared with all of the training data ， so that we have ten thousand scores to draw DET Curve.

Table 2: Introduction to Experiment Statistics

To draw a DET Curve, we have to prove that ten thousand scores are under same order of magnitudes. In GMM-SVM, all scores come from the distance of the testing vectors to the hyper plane, which is determined by the SVM training. Similarly，when we use UBM vector as a reference in verification regardless of the difference between the testing vectors and the training vectors. The scores we limit in 0-2. As a resul, we can demonstrate the curves in a same picture.

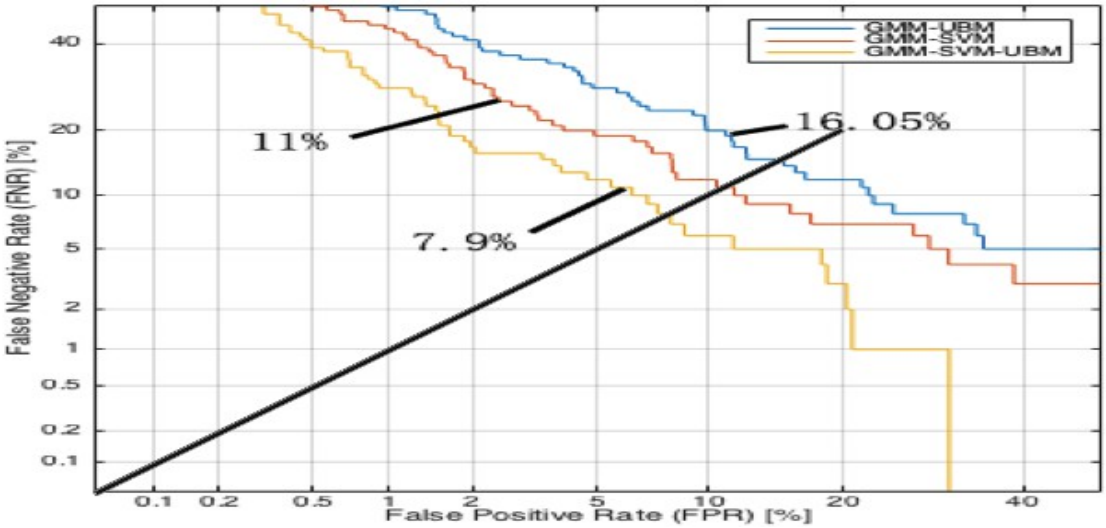

Figure 3: DET Curves of Three Schemes

We make function adapt into DET curve because of symmetry of the False Negative Rate（FNR）and the False Positive Rate（FPR）. In DET Curves,

FNR represents the possibility of a imposter who is be tested as the true speaker. FPR represents the possibility of a true speaker who fails to be recognized. Above all. Equal Error Rate（EER）is equal to the point of intersection of DET Curve and $y=x$.

It can be seen from Figure 3 that the GMM-SVM (EER: 11%) can be improved by nearly 4% on the basis of GMM-UBM (EER: 16.05%), indicating that SVM can be applied to the speaker recognition and features higher reliability. In the course of the experiment, the calculation time of GMM-SVM is 5.3h, much higher than the calculation time of GMM-UBM. As the speaker recognition focuses on the accuracy of the recognition, there is not much research on the computing time. Upon the improvement of the computer hardware, it can reduce the shortage of time.

If we use UBM vectoring as a stable reference（EER：7.9%）, the improvement of EER will be more remarkable that the system rises by nearly 8% because the area of speaker verification and vectoring the parameters will have the difference between speech data enlarged to show more characteristic of speakers. We also find that the computing time of GMM-SVM-UBM is only 2.3h， much less than that of GMM-SVM.

6.Summary

In this paper, the speech information is expressed by the high dimension vector and the package by using HTK voice format to avoid loss of speaker information to transform the recognition problem into a classification problem. Adding SVM to the text independent speaker recognition improves the reliability in comparison with GMM. This paper selects Linear Kernel Function and we can try to select other functions in future research with better results of recognition in theory. At meanwhile, with UBM vector as a reference, it is also a reliable schedule to take into consideration. In a word, vectoring Gauss Mixture Model mean parameters will reflect more differences between testing speech and training ones and we will devote ourselves to research other ways that could make the speaker verification more precise .

References

[1] D. A. Reynolds, T. F. Quatieri, and R. Dunn. *Speaker verification using adapted Gaussian mixture models*[J]. Dig. Signal Process. vol. 10(1-3):19-41(2000).

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. *Front-End Factor Analysis For Speaker Verification*[J]. IEEE Transactions on Audio, Speech and Language Processing, vol. 19(4): 788-798(2010).

[3] D. Reynolds, T. F. Quatieri, and R. B. Dunn. *Speaker verification using adapted gaussian mixture models*[J]. Digital Signal Process, vol. 10(7),19–41(2000).

[4] 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 Proc*[J]. IEEE Int. Conf. Acoust. Speech Signal Process. Toulouse, 1,97-100(2006)

[5]Weizhong Zhu,Douglas O.Shaughnessy. *Incorporating Frequency Masking Filtering in a Standard MFCC Feature Exteaction Algorithm*[J]. Proc.7th International Conference on Signal Processing,ICSP,617-620(2004).

[6] Steve Young, Gunnar Evermann, Mark Gales. *The HTK Book[EB/OL]*[Z]. http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml, (2006)