Research of the speaker verification based on the SVM-GMM mixture model

Xuan Cui Bo Deng Wen Zhuang

School of Mathematics & Computer Engineering, Xihua University, Chengdu 610039, P. R. China

Abstract

We put forward a new SVM-GMM mixture model to improve recognition rate of the speaker verification system in the paper. Support vector machines (SVM) and Gaussian mixture model (GMM) are widely applied to the speaker verification, but both have some disadvantages. We present a new approach for speaker verification based on their feature. The new model introduce the output of the Gaussian mixture model to Support vector machines, in order to adjust the probabilistic output of the support vector of machines. It can compliment support vector machines with probabilistic information. The experiments have proved that SVM-GMM mixture model can effective enhance the recognition rate of the speaker verification system.

Keywords: Speaker verification, Gaussian mixed model, Support vector machines, SVM-GMM mixture model

1. Instruction

With the developing of the science and technology, the information systems move forwards intellectualization. Therefore, speaker recognition become vigorously. It widespread application to many fields. Such as the negotiable securities transaction, automobile voice lock, National defense monitor and so on.

Speaker recognition, which can be classified into speaker identification and speaker verification. Feature extraction and establishing classification model are the basic problems in speaker identification and verification systems [1]. This article aimed at the latter problem to study, we present a new method to establish classification model and apply SVM-GMM mixture model to speaker verification. It can provide complimentary information to the Gaussian mixture model and the Support vector machines.

Following, the principle of the GMM and SVM are discussed. Latter, the new SVM-GMM mixture model is introduced. At Last, the experiment detail and conclusion are given.

2. Background knowledge

2.1. Gaussian mixture model (GMM)

At present, Gaussian mixture model (GMM) often to be used to the speaker recognition, this model has the good ability of recognition [2].

A GMM is a weighted sum of M component densities and is given by the form

$$p(X/\lambda) = \sum_{i=1}^{N} c_i b_i(x) \qquad (1)$$

where x is a dimensional random vector , $b_i(x)$, i = 1,. . ., N, is the component densities and c_i, i = 1,. . .,N, is the mixture weights. Gaussian function of the form

$$b_i(x) = \frac{1}{(2\pi)^{d/2}|\sum_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)\} \qquad (2)$$

with mean vector μ_i and covariance matrix $\sum_i$ The mixture weights satisfy the constraint that:

$$\sum_{i=1}^{N} c_i = 1 \qquad (3)$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation:

$\lambda = \{c_i, \mu_i, \sum_i\}$, i=1…N

In speaker recognition system, each speaker is represented by such a GMM and is referred to by this model λ.

For a sequence of T test vectors X = $x_1, x_2, \ldots, x_T$, the standard approach is to calculate the GMM likelihood in the log domain as:

$$L(X|\lambda) = \log(X|\lambda) = \sum_{i=1}^{T} \log(x_i|\lambda_i) \qquad (4)$$

The speaker-specific GMM parameters are estimated by the EM algorithm using training data uttered by the corresponding speaker using the HTK toolkit [1, 2, 5]

2.2. Support vector machines (SVM)

The following is a brief overview of Support vector machines (SVM), many details maybe found in Burges. SVMs work by constructing a binary classifier from a set of labeled points which form the training set. Let (x_i, z_j), i$\in$[1,2 ...,N]. be the training set where $x_i \in R^d$, is the d-dimension input feature vector and z $=\pm 1$ is the class labeling. The aim is to train a machine to learn the mapping x-->z_j. Such that the number of errors which are minimized. This is to be achieved using a function, f(x,α), where α is an adjustable parameter (or set of parameters) [3].

In the linearly separable case, the problem reduces to determining a hyperplane that divides the two classes. The two parameters w and b may always be rescaled such that

$$z_i[< w.x_i > +b]-1 \geq 0 \qquad \forall i \tag{5}$$

It can be shown that the optimal separating hyperplane (OSH), the hyperplane with the largest margin, can be obtained by minimizing w: This is a quadratic programming, usually solved using Lagrange multipliers.

In many instances the classes may not be linear, they are introduced to separable. Therefore, slack variables, ξ_i are introduced to allow for misclassifications, where ξ_i >0, thus the solution for the OSH is equivalent to minimizing

$< w, w >/2 + C\sum_i^N = 1\xi_i.C$ is a parameter chosen a priori by the user's controls. The capacity of the system, and is usually determined experimentally.

It is also possible, for instances where the classes are nonlinearly separable, to use kernel functions. The most common kernels include: linear kernels, polynomial kernels, radial basis functions (RBF) and multilayer perceptrons. Gaussian RBFs, as used here, are described by the following equation:

$$k(x, x_i) = \exp(1/2y^2 \parallel x - x_i \parallel^2) \tag{6}$$

where γ is also chosen a priori by user. Gaussian RBFs have been found to have a very good generalization performance [6]-[7]. The optimization problem is therefore the maximization of

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j z_i z_j k < x^T_{i} x_j > \tag{7}$$

subject to the constraints

$$0 \leq \alpha_i \leq C \tag{8}$$

$$\sum_{i=1}^{N} a_i y_i = 0 \tag{9}$$

At present, the linear kernel function, multinomial kernel function and the Gaussian kernel function are spread used in speaker recognition [6]. support vector machines (SVM) seek the optimal separating hyperplane (OSH), between the different classes reflect the difference between the different classes data, but its training time is long, and cannot reflect the characteristic .of training data .

3. Related Work

Traditional methods for the speaker recognition are GMMs, vector quantization, artificial neural networks, And SVMs [1, 4]. Of these methods The GMMs have been the most fashionable, because of many factors, including a probabilistic framework, and high-accuracy recognition; but usually, the difference of data, among varied classes have been easy to neglect .The SVMs also have been used more, the first set of approaches attempts to model emission probabilities for Hidden Markov Models . This approach has been moderately successful in reducing error rates, but suffers from several problems. First, large training sets result in long training time for support vector methods. Second, the emission probabilities must be approximate [4]. Third, the compute is very complex, since the output of the support vector machine is not probability. To overcome those shortages, as we described above, the new classification Model is generated.

4. SVM-GMM Mixture Model Establish and Analysis

Because SVM model and GMM model Each one has advantages and disadvantages, In this paper, we make use of their merits, establishes SVM and the GMM mixture model, Unifies the GMM logarithm according to the attribute ability with the SVM logarithm and the separating strong capacity characteristic, carries on the adjustment, introduces GMM to the SVM output to realize probability output of SVM.

4.1. SVM model introduction

Concrete algorithm as follows:

The supposition support vector machine output form is

$$y=\mathrm{sgn}(f(x)) \tag{10}$$

$$f(x)=(w.\ x)+b \tag{11}$$

where x is a input vector; w is a weight vector, b is a threshold value [6].

In computation process, the training sample has carried on the normalization, namely, the nearest spot to classified surface may satisfy |f(x)|=1, demand

$$\min|(w.\ x_i)+b|=1 \tag{12}$$

where x_1 ,x_2 …,x_n are in training data set [9].

In fact, the output of SVM is a distance Measure, we can project the distance of the SVM to posterior probability by used a mapping function. May produce the support vector machine output form through the Signoid function as follows:

$$p(\lambda \mid x) = p(y = 1 \mid x) = p(x) = \frac{1}{1 + \exp(-f(x))} \quad (13)$$

which is equal to:

$$p(C_{+1} \mid x) = \frac{1}{1 + \exp(-f(x))} \quad (14)$$

$$p(C_{-1} \mid x) = \frac{1}{1 + \exp(f(x))} \quad (15)$$

, where C_{+1} and C_{-1} means the sample set of +1 and -1 class, $C_{+1}=\{x_1,x_2\ldots\ldots x_m\}$, $C_{-1}=\{ x_1,x_2\ldots\ldots x_m \}$. Obviously, occupies on the classified surface point correspondence +1 and -1 kind of probability both are 0.5[5, 8]. Because equation (14)and (15) the output of the SVM, only can be affirmed by the distance between samples and the optimal separating hyperplane (OSH), it rarely reflects the distinction of the two classes, it can not reflects the distribution situation of samples which among the same class, therefore, the SVM has some limitation . In order to enable the model to realize the probability output, GMM is introduced.

4.2. GMM Introduction

In this section, we will elaborate introduce how to introducing GMM, and apply GMM to adjust the output of SVM,

Firstly, GMM is introduced in SVM model:

According to the 2.1 section, random vector x_i (i=1,2…,N) in the set of samples C_k (k=+1 or -1) , N is the number of the vector in the set of samples C_k [8, 9]. The output of the GMM as follows:

$$P_{GMM}(x_i \mid C_k) = \sum c_{km} N(x_i, \mu_{km}, \Sigma_{km}) \quad (16)$$

where

$$N(x_i, \mu_{km}, \Sigma_{km}) = \frac{\exp[-\frac{1}{2}(x_i - \mu_{km})^T \Sigma_{km}^{-1} (x_i - \mu_{km})]}{(2\pi^{d/2} \mid \Sigma_{km} \mid)^{1/2}} \quad (17)$$

$c_{km}, \mu_{km} and \Sigma_{km}$ respectively mean weight, average and covariance of m^th mixture for class k [1, 8].In equation (16), we Insert the GMM probability output into the SVM probability output, maker sure that the output of SVM, which not only considered the information in classes, but also considered the information among each class.

Following, we establish the SVM-GMM mixture model. In order to adjust the output of SVM by GMM, introduce the adjustment factor $S(x_i, C_{+1})$ and $S(x_i, C_{-1})$, the forms as follows:

$$S(x_i, C_{+1}) = \begin{Bmatrix} P_{GMM}(C_{+1} \mid x_i) + 0.5 & f(x) \geq 0 \\ 1.5 - P_{GMM}(C_{+1} \mid x_i) & f(x) < 0 \end{Bmatrix} \quad (18)$$

$$S(x_i, C_{-1}) = \begin{Bmatrix} 1.5 - P_{GMM}(C_{-1} \mid x_i) & f(x) \geq 0 \\ P_{GMM}(C_{-1} \mid x_i) + 0.5 & f(x) < 0 \end{Bmatrix} \quad (19)$$

Where

$$P_{GMM}(C_{+1} \mid x_i) = \frac{P_{GMM}(x_i \mid C_{+1})}{P_{GMM}(x_i \mid C_{+1}) + P_{GMM}(x_i \mid C_{-1})} \quad (20)$$

$$P_{GMM}(C_{-1} \mid x_i) = \frac{P_{GMM}(x_i \mid C_{-1})}{P_{GMM}(x_i \mid C_{+1}) + P_{GMM}(x_i \mid C_{-1})} \quad (21)$$

The two adjustment factors are used to adjust the probability output of the SVM model. $S(x_i, C_{+1})$ and $S(x_i, C_{-1})$ are fused in equation (14) and (15), get the SVM-GMM mixture model which we want to establish. Supposed input vector is x_i, take following format to Fuse, the two kinds of posterior probability as follows:

$$p(C_{+1} \mid x_i) = \frac{1}{1 + \exp[-f(x_i) S(x_i, C_{+1})]} \quad （22）$$

$$p(C_{-1} \mid x_i) = \frac{1}{1 + \exp[-f(x_i) S(x_i, C_{-1})]} \quad （23）$$

Obviously, this kind of model manifests that SVM and GMM have been well union in this paper. It make use of the adjustment factor to change variable of the SVM, for enhance the ability of the probability output.

Through this new model apply in our experiments, we can realize the adjustment. The detail process as follows. Make use of GMM to adjust the output of the SVM, when the result of the SVM is equal to the GMM, enhance the output of the SVM, conversely, when the result of the SVM is unequal to the GMM, decrease the output of the SVM.

In this Analysis show that introducing the classified result of GMM to the probability output of SVM, have an effect on adjustment, meanwhile the information between inner class and classes can incarnate by the new SVM-GMM mixture model.

5. Experiment

In the experiment uses the pronunciation data, which are 25 speakers, contains 10 female, 15 male, are distanced 30 day-long time intervals to record two pronunciations. The sound recording carries on under the laboratory environment, all is uses the 8K sampling rate. The speakers complete are students, each person reads aloud 1 to 9. The sentence length is 10s to 30s, each sentence saves is a document, each time enrolls 5 sections of sentences. Before the feature parameter extraction withdraws carry on preemphasis, use 1-0.9375 Z^{-1}, participle processing, each section of pronunciations division for frame long 30 ms, 240 sampling spots, the frame move is 0, the frame number is 48, add the Hamming window to be smooth the signal. In the experimental system selects two methods:

- Base on 16 steps MFCC voice feature parameter and GMM method to establish a classification model.
- Base on 16 steps MFCC voice feature parameter and the new SVM-GMM method to establish classification model.

In this paper we compare the speaker verification results based on our new SVM-GMM mixture model with current state-of-the-art SVM method.

We give the error rate contrast chart, and analysis the miss probability of the system in training time. The experimental data chart as follows:

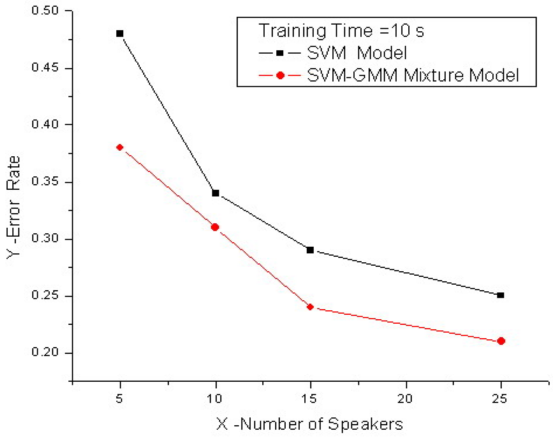

Fig. 1 Error Rate Contrast. in 10 Seconds.

It is shown that when training time is 10s, both SVM system and SVM-GMM system error rate are high, but with the SVM-GMM method is better than SVM method.

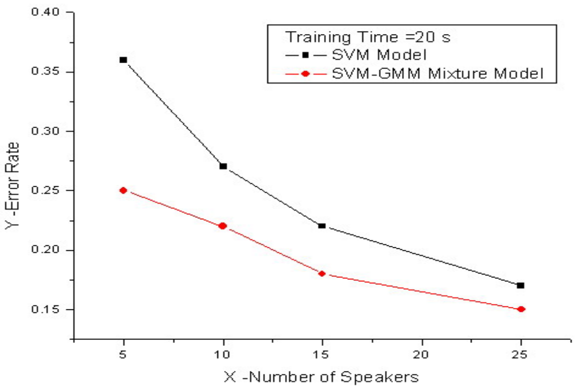

Fig. 2 Error Rate Contrast. in 20 Seconds

As it's shown that increase the training time, both miss probability of the SVM system and SVM-GMM system decrease, but the error rate of the SVM system is higher than SVM-GMM system.

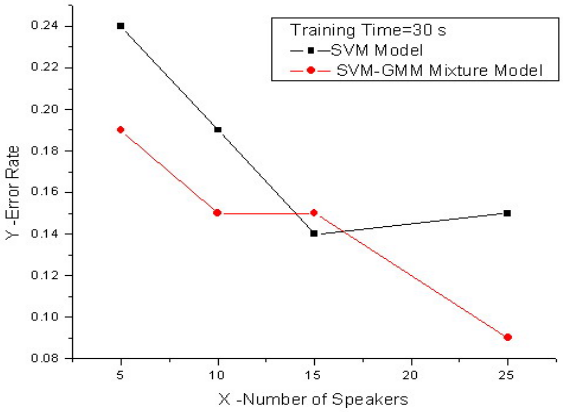

Fig. 3 Error Rate Contrast. in 30 Seconds.

It is shown that, the miss probability of the new method large scale drop when increase the training time ,the more speakers the more remarkable, and it have better effect on verification than traditional one.

Average rate verification of the SVM model system and SVM-GMM system are given, the experimental result as follows:

Training time/s	10	20	30
	Average Rate of recognition (%)		
5(speakers)	0.52	0.64	0.76
10(speakers)	0.66	0.73	0.81
15(speakers)	0.71	0.78	0.86
25(speakers)	0.75	0.83	0.85

Table1: Average Rate of SVM Method.

From Table1 we are easy to see that the rate of the recognition is improved by increasing the number of speakers, and it is also grown by increasing the training time, but the rate is small.

Training time/s	10	20	30
	Average Rate of recognition (%)		
5(speakers)	0.62	0.75	0.81
10(speakers)	0.69	0.78	0.85
15(speakers)	0.76	0.82	0.85
25(speakers)	0.79	0.85	0.91

Table2: Average Rate of SVM-GMM Method.

It is shown that the rate is improved by increase the number training time and number of the speakers.

We compare the speaker verification results based on our new SVM-GMM mixture model with the traditional SVM. we can see that when there are few or many speakers in our test ,the rate of the system which established by SVM-GMM mixture model are higher than traditional one, namely, the data matching of SVM-GMM mixture model is better than SVM model ,so this approach can improve the rate of the recognition . However, the results also show that with the increase number of speakers, the more speakers, the less improvement rate of recognition. Because of increasing the complexity in the new systems and for our collection set is too small and there is limited to the subject, the adaptability of our new model is still in need of further experiment.

6. Conclusions

In our new mixture model ,the most important merit is that Unifies the probability of GMM and the result of SVM, the new SVM-GMM mixture model retained the merit of SVM ,has the ability to strong decision, and also has manifested the GMM 's ability to probability expression data, and adjust the output of the SVM by the GMM training perior probability knowledge. The output not only considered the distance of samples but also considered samples distributed situation, so that the posterior probability which obtained from our experiments can reflect the actual situation. To sum up, in the experiment, it is proved that the new SVM-GMM mixture model is applied to our speaker verification system, can remarkably improve the result of the recognition.

References

[1] S. Fine, J. Navratil, R.A. Gopinath. Hybrid GMM/SVM Approach to Speaker Indentification.*Proc.ICASSP*, 2001.

[2] N. Malayath, H. Hermansky, S. Kajarekar, B. Yegnanarayana, Data-driven temporal filters and alternatives to GMM in speaker verification. *Digital Signal Processing*, 55–74, 2000.

[3] V. Vapnik, The Nature of Statistical Learning Theory. *Springer-Verlag, New York*, 2001.

[4] D.l Garcia-Romero, Julian Fierrez-Aguilar, Using quality measures for multilevel speaker recognition *Computer Speech and Language* 20:192–209, 2006.

[5] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process*. 3 (1):72–83, 1995.

[6] D. Xin, Z. Wu, Speaker recognition using continuous density support vector machines, *Electronics Letters*, Vol. 37, No. 17, pp. 1099-1101, 2001.

[7] D. Xin, Z H Wu. Speaker Recognition Using Continuous Density Support Vector Machines .*IEEE Electronics Letters*, 37(17):1009-1011,2001.

[8] S. Nakagawa, W. Zhang, M. Takahashi, text-independent/text- prompted speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM *IEICE Trans. Inform.Syst*. 89-D (3), 1058–1064. 2006.

[9] W.M. Campbell , J.P. Campbell, D.A. Reynolds E. Singer, P.A. Torres-Carrasquillo Support vector machines for speaker and language recognition *Computer Speech and Language* 20:210–229, 2006.