IMPLEMENTATION OF A SPEAKER VERIFICATION SYSTEM THROUGH COMBINATION OF GMMS AND SVM AND LINEAR COMBINATION OF GMM AND SVM

Azar Mahmoodzadeh
Department of Electrical Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.
E-mail: azar_mahmoodzadeh@yahoo.com
Razieh Darang
Department of Electrical Engineering, Bushehr Branch, Islamic Azad University, Bushehr, Iran.
E-mail: raz.darang@yahoo.com
Babak Ranjbar
Department of Electrical Engineering, Delvar Branch, Islamic Azad University, Delvar, Iran
E-mail: Ranjbar.babak@gmail.com
Nooshin Rabiee*
Young Researchers and Elite Club, Bushehr Branch, Islamic Azad University, Bushehr, Iran.
*Corresponding author: E-mail: nooshin.rabiee@yahoo.com, Phone: +989177019794
Ghasem Ravaee
Department of Electrical Engineering,Kazeroon Branch, Islamic Azad University,Kazeroon ,Iran
E-mail:Gravaei58@gmail.com
Maryam Rahimi Kazerooni
Department of Electrical Engineering, Kazeroon Branch, Islamic Azad University,Kazeroon ,Iran.
: -mail: Mar.rahimi63@gmail.com

ABSTRACT : *In this paper, a speaker verification system is implemented in a text-independent procedure. The methods employed for implementation include the Gaussian mixture models, support vector machine and linear/series combinations of Gaussian mixture model and support vector machine.*

In the series combination of the two classifiers, GMM is the base model while SVM is applied as a post-processing for curtailing the classification error. That is because the Gaussian mixture model is not adequately capable for classification in cases where the data are extremely approximate and there will be an extreme error rate. In the linear similarly, since both PSVM and GMM are inherently capable in speaker verification, each one is trained individually to derive an output linear combination for verifying the speaker's identity claim, in which scenario better results will be achieved.

Keywords: Text-independent speaker verification, Gaussian mixture model, Support vector machine, Probability support vector machine, Half total error rate.

1. INTRODUCTION

Each of the productive and distinction models entails features in which others perform in a weaker way. For example, producing models are less capable in separating approximate data unable to be separated linearly, thus curtailing the performance of speaker verification systems. In contrast, the distinction models take advantage of certain capabilities for blending the data and yield a classification at an acceptable error rate. That is why most speaker verification and recognition systems today attempt to decrease error rate through combination of productive and distinction methods [3,8, 11, 13, 14].

Speaker recognition algorithms have numerous applications such as speaker verification, individual access to database through phones, bilingual interpretation etc. Effort has been made so far in practical research concerning speech processing to improve the performance of such algorithms in order to achieve high accuracy and reliability in speaker recognition [6.10].

Speaker recognition can be divided into two types of speaker recognition or identification and speaker verification. A person utters a sentence, the goal is to recognize the speaker and tell which M authorized speaker it is. If the goal is to verify the identity a person claims, the verification system will either approve or reject. Each of these realms can in turn take place as text-dependent and test-independent. The former implies that recognition requires the person to utter a certain sequence of words, so the system could identify the speaker, whereas the former is capable of recognize the speaker regardless of what the person utters [2].

2. Gaussian Mixture Model (GMM)

Gaussian mixture model is the most widely used approach in the field of speech recognition and speaker recognition. In fact, the Gaussian mixture model estimates the density written as a sum of several Gaussian functions. If *m* is the number of Gaussian mixtures and *D* is the feature vector dimensions, then:

$$P(x|M)=\sum_{i=1}^{m} a_i \frac{1}{(2\pi)^{D/2}\sum_i^{1/2}} \exp(-1/2(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)) \quad (1)$$

Where $P(X|M)$ represents the probability of x belonging to the speaker *M* equivalent to the total *M* Gaussian functions with μ_i means and covariance matrix Σi. The parameters corresponding to each mixture is estimated through using the algorithm EM[5] [7].

The Gaussian mixture model can be regarded as an extension to vector quantization, i.e. the mean of each Gaussian mixture can be considered a code book. The Gaussian mixture model approach blends the parametric Gaussian density and non-parametric vector quantization. Normally, the GMMs are applied along with diagonal covariance matrix for the convenience of calculation [1].

If *X* is the test term with *N* frames, then rating calculated for each model shall be as follows:

$$S(x)=\log p(x|M)=1/N\sum_{i=1}^{N} \log p(x_i|M) \quad (2)$$

Gaussian mixture model is the most widely used method for speaker identification and verification. This model actually estimates the density written in the form of several Gaussian functions. Selection of an appropriate density function depends on the feature vector and the intended application. In the text-independent speaker verification where there is no prior knowledge of what is being uttered, the Gaussian mixture model has been remarkably successful [5, 12].

As for the text-independent speaker verification, the Gaussian mixture model entails an inherent talent for successful classification of data regarded as the most widely used classifier in this respect, having turned into a successful classic method over the last two decades.

3. Linear classification

If there is a linear function that can classify information without error (i.e. the information points belonging to a class be placed on one side of the line or plane or hyper-plane), the information will be known as linearly capable of separation. In cases where the information can be separated linearly or can be separated linearly at a good approximation, the linear classifiers will be applied. The advantage of such classifiers lies in their simplicity and easy implementation. This classifier can be indexed through the following form:

This classification can be used to index the form below:

$F(x)=(w.x)+b \quad w€ R_N \quad ,b€ R$ (3)

Where *W* is an *N*-dimensional vector. When the data dimensions are N=2, the left side of the above equation adopt zero representing a plane. When the When the data dimensions are N>2, it will represent a hyper-plane, the position and orientation of which are specified through adjustment of a set of a=[w b] parameters. In equation (3), *w* is the normal vector of the hyper-plane.

4. Nonlinear classification

In many cases, the data are not linearly separable, i.e. the information has been laid out in a way there is no linear function found to be separated error-free through line, plane or hyper-plane, because the classification error through a linear classifier may be too undesirable. Due to the use of a kernel function, the support vector machine can act as a non-linear classifier. Hence, the application of an appropriate kernel function can improve to a considerable extent this classifier against other linear classifiers. In [15], the researchers showed how a support vector machine output can be a polynomial classifier through a polynomial kernel function.

5. Mathematical analysis of support vector machine

In SVM, each data is seen as a P-dimensional vector (or a list of *P* number). It is intended to figure out is such points can be separated through a P-1-dimensional hyper-plane, which is called a linear separation. There are numerous hyper-planes capable of separating the data. The question is what hyper-plane to choose, the concept of training data classifiable as points in a high-dimensional space and the line for separation are not unique. What distinguishes the SVM from other separators is the hyper-plane selection procedure.

In SVM, the goal is to maximize the margin between two classes. Hence, it chooses a hyper-plane where the distance from the closest data on both sides of the separator is a maximum line. If there is such a hyper-plane, it will be identified as a maximum-margin hyper-plane [9, 16]. Figure 1 illustrates this concept visually.

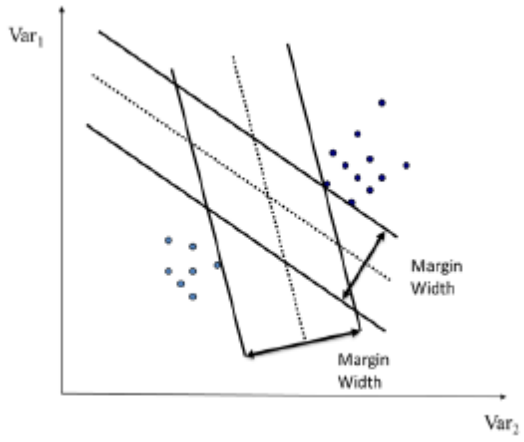

Fig. 1 Depiction of hyper-planes for selection of the maximum-margin hyper-plane.

The reasons why a maximum-margin hyper-plane needs to be selected have been given below:

- It logically seems to the safest strategy, the farther the margin surrounding the separator line from the data the greater the chances of correct recognition for the training data.
- There are theories based on the VC dimension proving its usefulness.
- This approach has empirically worked very well.

In order to build the maximum margin, two boundary planes are drawn parallel to the separator plane at two farthest points to collide with the data. The plane with the farthest distance from the boundary plane shall be the best separator, while the decision-making boundary us actually perpendicular to the line connecting the boundary planes [4, 6].

As for data separation, the decision function is determined through a subset of training vectors (closest to the hyper-plane).

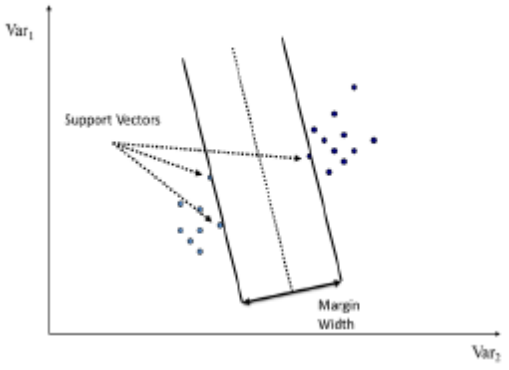

Fig. 2 Optimized hyper-plane and support vectors. In fact, the optimal hyper-plane in SVM is the separator between the support vectors.

The proper use of SVM will help the algorithm yield a good generalization. Despite the large size, it avoids the overflow. Moreover, due to the use of support vectors instead of the entire data, the algorithm also involves data compression.

6. Probability support vector machine (PSVM)

In practical Pattern Recognition issues, it will be beneficial to obtain the posterior probability of an entry membership in a particular class. For example, the posterior probability demonstrates its power when the classifier is a small portion of a larger decision-making, employing a Bayesian optimality criterion. Having understood the importance of obtaining the posterior probability in classifiers, it is essential to obtain the same in the SVM involved as a score in the main classifier so as to improve the results. The support vector machine produces an inaccurate value which is not the probability value. If the SVM output is:

$f(x)=h(x)+b$ (4)

While:

$H(x)=\sum y_1 a_1 k(x_1, x)$ (5)

Then, the core of equation (4) should be minimized in the Hilbert space:

$c\sum(1-y_1 f_1)+1/2 \| h_F \|$ (6)

The potential output production procedure in a core machine has been proposed in [3]. Using a logical function, we have the following equation:

$P(CLASS \mid INPUT)=P(y=1 \mid x)=p(x)=1/1+exp(-f(x))$ (7)

7. Methods of combining Gaussian mixture model and support vector machine

In the Pattern Recognition science, there are different ways to combine classifiers. This kind of combination originates from the nature of classifiers and their compositional properties.

Concerning the two combination methods, the following three main ideas can be mentioned:

1. Using an output combination for both models either linearly or non-linearly for the trained models separately
2. Using the output of either model, usually the productive model, for selection of training data in the other model.
3. Combination to create new formulations for minimizing the error rate simultaneously in productive modeling and distinction modeling

In this article, the first two methods have been used for combination of the classifiers. In the former, the output line of the GMM and PSVM classifiers were combined through predefined coefficients.

In the latter, certain cases involved the GMM output of to selection of training data in the SVM upon which the decision-making was assigned. Later on, both methods will be explained.

8. Linear combination of Gaussian mixture model and support vector machine

The basic idea of combining the SVM GMM classifiers is very simple, because they could independently in the previous research demonstrate their inherent ability. It can be perceived that linear combination of outputs from the two models can be useful. A conventional method for multi-classifier combination involves the linear combination of posterior probability in an entry membership with a particular Class (P (class ׀ input)). The output of a Gaussian mixture model is made of probability lying inherently within the interval [0, 1]. As previously mentioned, the output of a PSVM model is also made of probability lying within the interval [0, 1]. Hence, the can be easily incorporated through coefficients (1 -λ) and (λ).

Achieved by trial and error, λ is a coefficient chosen so that the best output possible is obtained at the minimal error. The calculation formula of linear combination is as follows:

If P_{PSVM} is the score for PSVM model and P_{GMM} is the score for the GMM model, then:

$P_{PSVM} .P_{GMM} +(1-\lambda)P=\lambda$ (8)

9. Combination of Gaussian mixture models and support vector machine

In the series combination of these classifiers, the Gaussian mixture model is regarded as the base model while the support vector machine acts as the post-processing for error reduction. As such, in cases where the claiming speaker's score is within the upper threshold (t) and bottom (1-t) (the *t* threshold refers to the level calculated for each speaker in the GMM), where the GMM might mistakenly reject or approve it, the decision-making is assigned to the SVM.

Assisted by the Gaussian mixture model output data failing to made a good decision on the classifier, the support vector machine is trained (i.e. training whose output score lies between the upper and bottom thresholds as compared to the GMM) and decision is made on their identity claims. In this scenario, decision is made at a lower error rate.

10. Combination of GMM and SVM

At first, the Gaussian mixture models concerning the target speakers are trained through the training data. Then, the entire training data are divided into 1.5-second speech fragments the score of which is obtained against the GMM of each speaker. The core is compared with two threshold levels t and 1-t. The feature vectors for the fragments whose score compared to the speaker model lies between t and 1-t were applied for training the SVM of the same speaker. The t threshold level for each speaker was the same that calculated in the first experiment for each speaker in the GMMs.

In the testing phase, the speech fragment is first tested against the GMM corresponding to a given speaker whose score is calculated. If the score is higher than t, the speaker will be verified, and if it is less than 1-t, the speaker will be rejected. In case the score is somewhere between t and 1-t, the speech fragment will be assessed through the SVM of the same speaker, where the SVM decides whether or not the speaker is verified.

The table below shows the amount of errors "false approval" and "false rejection" as well as the HTER criteria for the number of different target speakers in verification using combination of GMM and SVM models. The training and test data are similar to previous experiments and the number of GMM Gaussians is 64. Figure 3 displays the HTER for a number of different speakers whose number increases at higher error rates.

11. Linear combination of GMM and PSVM

An alternative method involves the linear combination of GMM and PSVM outputs. In this scenario, the GMM and PSVM are individually trained for each target speaker. In the test phase, the score of each model is calculated for the input speech fragment. Each of these scores will adopt a value between one and zero. At the next stage, these scores are linearly combined according to the following formula.

$S=\lambda s_{GMM} + (1-\lambda)S_{PSVM}$ (9)

Table. 1 The error rates for "false approval" and "false rejection" as well as the HTER criteria for the number of different target speakers in verification using combination of GMM and SVM models.

Number of	P_{FA}	P_{FR} (%)	HTER (%)
10	10.6	7.8	9.2
15	12.3	8.6	10.5
20	11.3	13.4	12.4
25	16.4	11	13.7
30	12.8	17.6	15.2
35	15	18.1	16.6

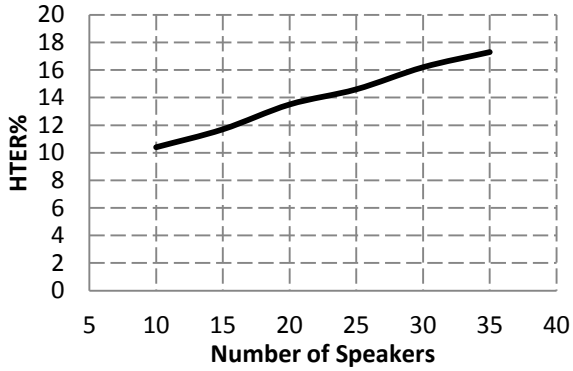

Fig. 3 The HTER criterion according to the number of different speakers in the speaker verification system using combination of GMM and SVM

Where S_{GMM} is the score for the Gaussian mixture model and S_{PSVM} is the score for the PSVM. The score obtained through linear combination will be compared with the threshold level, which is in this case obtained from the linear combination of threshold levels corresponding to the GMM and PSVM. If the score is above the threshold, the speaker will be verified, otherwise there will be rejection. The value of λ can be experimentally determined by trial and error. In these experiments, λ=0.6 was considered.

Table 2 displays the error rates for "false approval" and "false rejection" as well as the HTER criteria for the number of different target speakers in verification using combination of GMM and SVM models. The training and test data are similar to previous experiments and the number of GMM Gaussians is 64.

Figure 4 displays the HTER for a number of different speakers whose number increases at higher error rates.

Table 2. The error rates for "false approval" and "false rejection" as well as the HTER criteria for the number of different target speakers in verification using combination of GMM and SVM models

Number of target	PFA	PFR (%)	HTER (%)
10	7.2	9.5	8.4
15	12.3	7.6	10
20	13	10.6	11.8
25	11.5	14.8	13.2
30	17.7	12.3	15
35	13.7	18	15.9

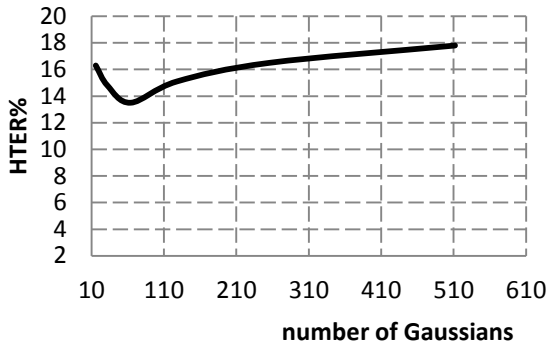

Fig. 4 The HTER criterion according to the number of different speakers in the speaker verification system using linear combination of GMM and SVM.

Similarly in Figure (5), the HTER can be observed and compared for the number of speakers based on the four categories in speaker recognition applied in this research. The training and testing data were similar to the previous version and the number of GMM Gaussians were selected to be 64

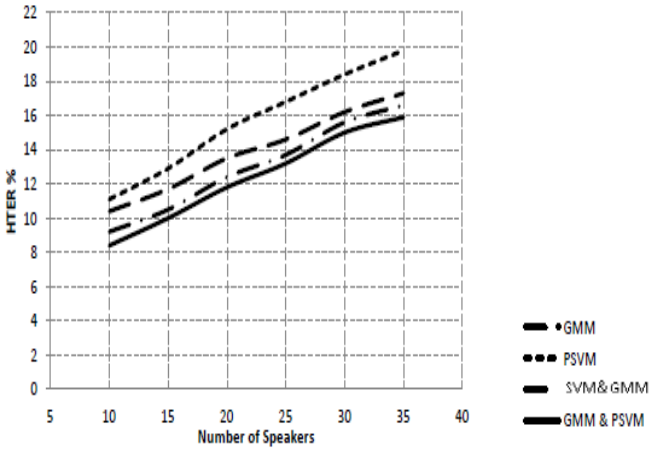

Fig.5 The HTER criterion according to the number of different speakers in the speaker verification system using series and linear combination of GMM and SVM

Figure 6 similarly displays the detection error tradeoff (DET) for various methods of speaker verification. The number of target speakers is 20 and the number of GMM Gaussians is 64. These curves are obtained for different values of the threshold level. As mentioned earlier, the threshold level is selected in a way to cover *p* percent of non-target speaker scores. In this graph, as the p value varies from 50 to 100 percent, the FA and FR errors are calculated and then the FA is drawn based on the FR. As can be seen in the curve, the GMM is more efficient than the PSVM. In addition, the combination of GMM and SVM will yield even higher efficiency as compared to either method individually. In terms of efficient, the linear combination is more ideal.

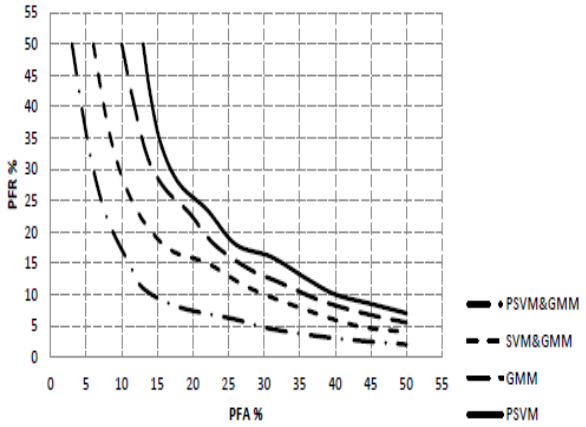

Fig. 6 The detection error tradeoff (DET) in speaker verification for the GMM and PSVM and their linear and series combinations.

12. Combination with other models for enhancing the distinction capability

The series combination of two classifiers involves the GMM as the base model while the SVM is applied as a post-processing error reduction, because the GMM entails training data extremely approximate, it is not adequately capable of classification and there will be a high error rate. In the linear combination, since both PSVM and GMM methods are inherently capable of accurate speaker verification, each can be individually trained and a linear combination of them be applied for speaker recognition. In this case, better results shall be obtained.

Comparing the results of implementing the above methods suggests that in combined classifiers, the HTER alone was curtailed as compared to the cases where one classifier was applied.

In scenarios where one classifier was applied, the GMM showed lower HTER than the PSVM. Similarly in application of classifier combination, the linear combination of GMM and PSVM was more efficient than the series combination of GMM and SVM in terms of the HTER.

13. CONCLUSION

In this article, the different methods for speaker verification were tested and evaluated as the results of experiments were discussed in terms of the number of different target speakers. The assessed methods included the Gaussian mixture model (GMM), probability support vector machine method (PSVM), series combination of GMM and PSVM, and linear combination of GMM and PSVM/ The various experiments revealed that combination of GMM and PSVM will yield the highest efficiency in speaker recognition at mean error rate of 11.8% for 20 speakers.

REFERENCES

[1] Sultani A. "Speaker Identification in a Telephone Conversation", Master thesis in Computer Engineering, Computer Engineering Department, Sharif University of Technology, 2010.

[2] Khayam-zadeh M. "Text-independent Speaker Identification through Gaussian Mixture Model" Master Thesis in Power Engineering, Department of Electrical Engineering, Amirkabir University of Technology, 2002.

[3] Nazari M. "Improvement of Phonemes Recognition in Continuous Speech through GMM and SVM", Master Thesis in Telecommunications, Department of Electrical Engineering, Amirkabir University of Technology, 2008

[4] A.Smola,b.sch"o1kopf,and K.M"uller.General cost functions for support vector regression.*In Australian Congress on Networks*,1998.

[5] Bimbot F.,et-al,"A Tutorial on Text-Independent Verification," *EURASIP Journal on Applied Signal Processing*,N.4,pp.430-451,2004.

[6] C.J.C.Burges.A tutorial on support vector machines for pattern recognition Data mining and knowledge Discovery,2(2):1.47,1998.

[7] F.Prenkopf,D.Bouchaf fra, "Genetic Based EM Algorithm for learning Gaussian mixture models "*IEEE Transaction on pattern Analysis and machine intelligence*. **Volume 27**,Issue 8(August 2005)pages:1344:1348,2005.

[8]G.Flaker,Heuristics for improving the performance of online SVM training algorithms. *In proc,NIPS 99*,1999.

[9] H.W.Kuhn and A.W.Tucker.Nonlinear programming.*In Proc.2nd Berkeley symposium on Mathematical Statistics and Probabilities*, pages 481.492,Berkeley,.University of California Press,1951.

[10] M.A.Przybocki.A.F.Martin,and A.N.Le,"NIST Speaker recognition evaluation utilizing the mixer corpora-2004,2005,2006,"*IEEE Transactions on Audio,speech,and Language processing* ,**Vol,15**.no,7,pp.1951-1959,2007.

[11] P.Clarkson and P.J.Moreno.On the use of support vector machines for phonetic classification.*In proc.ICASSEP*,pages 585.588,1999.

[12] Reynolds,Douglas A.,and Rose,R.C.,"Robust Text-Independent Speaker Identification Using Gussian Mixture Speaker Models,"*IEEE Trans.speech and Audio processing*,**Vol.3**,No.1,pp.72-83,1995.

[13] S.Fine,J.Naveratil,and R.A.Gopinath.A hybrid GMM/SVM approach to speaker identification.*In proc.ICASSEP*,2001.

[14] U.Kerebel.pairwise classification and support vector machines.In B.Scholkopf,C.Burges, and A.Smola,editors,Advances in kernel methods,support vector learning.*MIT press*,1999.

[15] V.Wan and W.M.Campbell.support vector machines for speaker verification and identification .*In proc.Neural Networks for signal processing* ,pages 775.784,2000.

[16]W.Karush.Minima of function of several variables with inequalities as side constraints.Master's thesis,Department of Mathematics,University of Chicago,1939.