**Yuhuan ZHOU[1], Xiongwei ZHANG[2], Jinming WANG[1], Yong GONG[2,] Yi ZHOU[3]**

Institute of Communications Engineering, PLA Univ. of Sci. & Tech. (1), Institute of Command Automation, PLA Univ. of Sci. & Tech. (2), Hydrometeorological Center of the South China Sea Navy(3).

# Speaker recognition based on the combination of GMM and SVDD

***Abstract***. *Score-level combination of subsystems can yield significant performance gains over individual subsystems in speaker recognition. A novel speaker verification method based on support vector data description (SVDD) is proposed to remedy the defect of Gaussian mixture model (GMM) to some extent, and then using the theory of multiple classifier systems (MCS),a new speaker recognition system based on the combination of GMM and SVDD is proposed. Experiments on TIMIT speech database show that the GMM-SVDD model fully utilizes the complementarities of GMM and SVDD to improve the performance obviously in speaker verification, closed-set speaker identification and speaker recognition.*

***Streszczenie.*** *Zaproponowano nowa metodę rozpoznawania głosu bazującą na systemie SVDD jako alternatywę dla modelu GMM. Następnie wykorzystując teorię wielokrotnego systemu klasyfikacji MCS zaproponowano wykorzystanie połączenia systemów GMM i SVDD. Eksperymenty potwierdziły że nowy model GMM-SVDD umożliwia ulepszonę rozpoznawanie głosu. (**Rozpoznawanie głosu bazujące na kombinacji systemów GMM i SVDD**)*

**Keywords:** Gaussian Mixture Model, Support Vector Data Description, Speaker Recognition, Multiple Classifier Systems
**Słowa kluczowe: r**ozpoznawanie głosu, systemy SVDD I GMM.

## 1 Introduction

Speaker recognition (SR) deals with identification of the person who utters a sentence and identity verification of the selected speaker using his/her voice. Two modes of operation are usually envisaged: speaker identification (SI) and speaker verification (SV). SI is concerned with the selection, while SV is concerned with the validation of the claimed identity of a person.

One-class classification is usually applied in SR, which can be divided into three categories: density estimation methods, reconstruction methods and boundary methods. Density estimation methods aim at estimating the whole distribution of the target data, typical methods are Parzen density estimation and GMM. GMM[1] has made great success in SR, especially in the closed set. However, when in open set SI or SV, the accuracy declines obviously because of instable likelihood score. The goal of reconstruction methods is to develop a simplified representation of the data via clusters or principal components, these methods are numerous: k-means, principal components analysis, self-organizing maps, etc. Boundary methods, rather than estimating the distribution, aim at constructing a boundary around the target data; K-nearest neighbours and SVDD are used.

SVDD is one of non-parametric models, which was originally suggested by Vapnik and interpreted as a novelty detector by Tax and Duin[2]. Usually, SVDD is used widely in pattern recognition, such as face recognition; however, its application in SR need more research. The purpose of SVDD is to given a compact description of the target data with a hyper-spherical model, which is determined by a small portion of data called support vectors. The boundary of each SVDD models could adjust easily by two important parameters: kernel parameter and penalty parameter, hence, the optimization verification threshold of SVDD could be found more conveniently than GMM.

In the last decade, many successful SR systems have relied on multiple classifier systems[3] to achieve superior performance. The motivation behind MCS is the fact that the response to the same input signal could be classifier dependent so the error of a given classifier could be corrected by the whole system. In MCS, classifiers are usually combined in parallel into two levels: abstract level and score level. In the abstract level approach, the binary decisions made by multiple classifiers are combined. In the score level configuration, the scores of individual classifiers are combined by mean of a set of weights in the last stage to produce the final score that is later threshold to obtain a decision. Intuitively, what we wish is to have individual classifiers that are highly uncorrelated. This way, all classifiers contribute independent and therefore, hopefully, complementary information leading to a better final decision[4].

In this paper, we consider the task of text-independent SR, as measured in the voice database of TIMIT, via score-level combination of GMM and SVDD. Firstly, a new text-independent speaker verification system based on a discriminatively trained SVDD classifier is proposed, which changed the crisp characterization to fuzzy characterization. Then, a new GMM-SVDD speaker recognition system fused SVDD and GMM is proposed, which relied on multiple classifier systems and adopt score-level combination method. Experiment results show that the accuracy rate of SI with SVDD is lower than GMM, but it obviously improves the accuracy and stability of SV system and the fusion model fully utilizes the complementarity of SVDD and GMM to improve the robustness and recognition accuracy in SR, SI and SV. When applied to a state-of-the-art GMM-SVDD model, this approach achieves average improvements of up to 37% in EER over baseline in SV and 27% in SR, and 33% of error rate in closed-set SI.

The structure of this paper is organized as follows. In Section II, we propose a new SR method based on SVDD, and a new GMM-SVDD speaker recognition system combined by GMM and SVDD. In Section III, we apply the proposed system to TIMIT database, and compare with baseline system. Concluding remarks and further research are presented in Section IV.

## 2 GMM-SVDD speaker recognition system

In this section, a new GMM-SVDD speaker recognition system fused GMM and SVDD is proposed. SR system contains two phases: training and recognition. In training phase, SVDD and GMM are constructed respectively. The difference between traditional model and GMM-SVDD is in recognition phase. Test speech will input the two models and the scores will be combined in the last stage to produce the final score. Specific processes are shown in figure 1, although the model structures of identification stage and verification stage are the same, the combination coefficients are different.

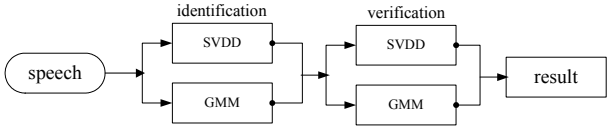


Fig.1. Flow chart of SR system fused GMM and SVDD

Let $G_{score}$ be the score of GMM, and $S_{score}$ be the score of SVDD, and then the final score of GMM-SVDD is:

$$F_{score} = \omega S_{score} + (1-\omega) G_{score}, 0 \le \omega \le 1 \quad (1)$$

Where $\omega$ is the combination coefficient. The proportion of the two models in final score is regulated by $\omega$. When $\omega$ =1, it represents that the final score was obtained by SVDD, and when $\omega$ =0, only GMM is used. $\omega$ will be changed from 0 to 1 to find the best combination system. The GMM-SVDD speaker recognition system has two key parts: SVDD and score normalization[5].

2.1 Support vector data description

SVDD was inspired by support vector machines (SVM), whose basic idea is to construct a hypersphere *(a,R)* with minimum volume containing most of the target data, where *a* is center, and *R* is radius of the minimum enclosing hypersphere, and the points in the sphere are the support vectors. The optimization problem is then given by:

(2) Minimize: $L = \sum_i \alpha_i K(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$

(3) Subject to $\begin{cases} \sum_i \alpha_i = 1 \\ 0 \le \alpha_i \le C, \forall i \end{cases}$

where $x_i$ are the support vectors and the center $a = \sum_i \alpha_i x_i$. In function (2) inner product operation $(x_i, x_j)$ is replaced by $K(x_i, x_j)$, which was called kernel trick, to find a more flexible data description in a high dimensional feature space.

Suppose *z* for the test samples, when meet (4), *z* is a target, otherwise an outlier.

$$f_{SVDD}(z,a,R) = I(\| \phi(z) - \phi(a) \|^2 \le R^2) = I\left( K(z,z) - 2\sum_i \alpha_i K(z,x_i) + \sum_{i,j} \alpha_i \alpha_j K(x_i,x_j) \le R^2 \right) \quad (4)$$

SVDD has several benefits[6]: first, it can model arbitrarily distributions without the assumption that the data are Gaussian distributed; second, fewer training samples are needed to train the models in high-dimensional spaces; third, the SVDD model could be trained only by one class data if there is no outlier data. Furthermore, the boundary of each SVDD models could adjust easily by two important parameters: kernel parameter and penalty parameter, hence, the optimization verification threshold could be found conveniently.

Because natural speech utterances carry information not only about who says it, but also about what is being said, emotional state, language species, and so on. It makes the distribution of sample points overlap in feature space, so the conventional SVDD doesn't work if only utilize *R* be crisp characterization. Actually, there is no sharp distinction between two different utterances, so traditional SVDD need improved to fit for SR[7].

A new fuzzy decision method is designed. Suppose a test voice $S_{test}$, whose total number of samples is *N*. For single sample recognition, we still use traditional methods. Suppose the number of accepted samples is *S*, then the likelihood score of $S_{test}$ is defined: $R_{score}=S/N$, where $R_{score} \in [0,1]$. If $R_{score}=1$, it means $S_{test}$ belongs to the model; while $R_{score}=0$, it means $S_{test}$ doesn't belong to the model, and when $R_{score} \in (0,1)$, it means $S_{test}$ belongs to the model for a certain extent. Therefore, the verification threshold *T* could be set between 0 and 1 according to the target acceptance rate and outlier rejection rate. Then the condition for the final decision is:

$$S_{test} = \begin{cases} 1, if R_{score} > T \\ 0, if R_{score} \le T \end{cases} \quad (5)$$

2.2 Score normalization

Score normalization strategies are used for reducing the variation in likelihood ratio scores of the fusion system in making SV decisions.

The use of score normalization techniques has become important in GMM based SV systems for reducing the effects of the many sources of statistical variability associated with log likelihood ratio scores. The sources of this variability are thought to include changes in the acoustic environment and communications channel as well as intra-speaker variation that may occur across multiple sessions.

For a given target speaker $s$ and a test utterance $u_{test}$, then the score of $u_{test}$ is $S(s, u_{test})$, and the normalized score is:

$$S_{norm}(s, u_{test}) = [S(s, u_{test}) - \mu] / \sigma \quad (6)$$

where $\mu$ is the mean and $\sigma$ is the standard deviation which need to be estimated. There are two well know score normalization methods: Z-norm and T-norm[8]. The different between the two methods are the computation of $\mu$ and $\sigma$. In the Z-norm, the parameters $\mu$ and $\sigma$ are estimated as the sample mean and standard deviation of a set of impostor speaker utterances. This represents an average of scores obtained by scoring the target speaker model against a set of impostor utterances. In the T-norm, the parameters $\mu$ and $\sigma$ are estimated as the sample mean and standard deviation of impostor speaker models. This represents an average of scores obtained by scoring a set of impostor speaker models against the test utterance. In SR, the first step is computing the scores of every speaker models to identify the most possibility speaker, so T-norm do not add any other computation. In this paper, score normalization method adopts the T-norm both in GMM and SVDD.

Suppose there are *N* different speaker models in a speaker set, expressed by $M_1, M_2 \cdots M_N$ respectively. For a test speech *V*, the initial score of each model can be obtained: $S_1, S_2 \cdots S_N$ in SI phase. Then in SV phase, the normalized score computed in actual is:

$$S_{norm} = S_I - \ln[\sum_{i=1}^{N} \exp(S_i)] \quad (7)$$

where *I* is the index of the most possibility speaker model. In this paper, the initial score $S_i$ is the fusion score $F_{score}$mentioned in function (1).

**3. Simulation**

This section presents an evaluation of the GMM-SVDD speaker recognition system performance under three scenarios: SV, closed-set SI and SR. Voice database used in the experiments is the TIMIT. 50 people selected in SI and SV experiment, while 100 people selected in SR. Each person has 20 speeches and each speech is about 2

seconds. Training and recognizing are selected 10 different speeches respectively. Frame length and frame shift are both 16ms, and 13-dimensional MFCC and 16-dimensional LPC are calculated from each frame.

3.1 Speaker verification

In SV experiment, the N-fold cross validation method is introduced to predict the average EER of each GMM-SVDD model, where *N* is 50. Each time select one speaker's voice data as target sample and the remaining 49 as outlier samples. The final system evaluation result derives from the average EER of 50 experiments.

Figure 2 shows the EER curve of the GMM-SVDD model with the combination coefficient $\omega$ changed from 0.8 to 1. The training speech of each model is 20 seconds.

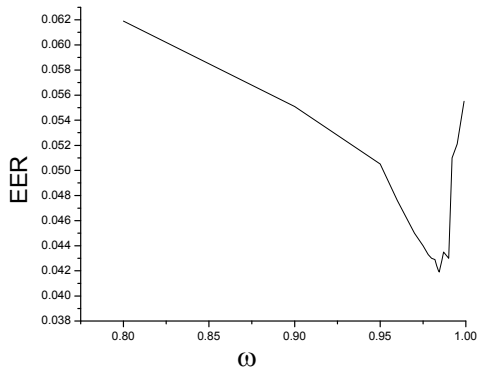


Fig 2. EER curve of GMM-SVDD changed with the combination coefficient

We can see that when $\omega$ =0.985, the EER of the GMM-SVDD reaches a minimum: 0.0419. When $\omega$ =1, only SVDD used, the EER is 0.0571, and when $\omega$ =0, only GMM used, the EER is 0.0739. Figure 3 shows the ROC curve of three models. It can be seen that the GMM results are significantly worse than the SVDD results, and the combination of GMM and SVDD systems leads to further improvements.

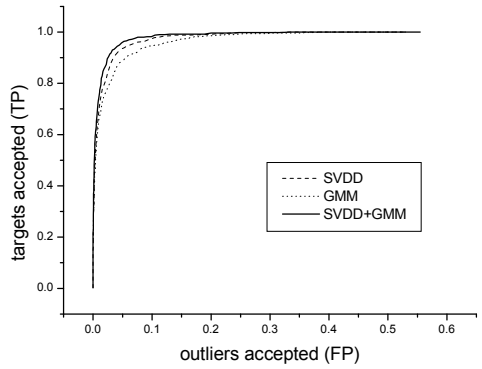


Fig3. ROC curve of three different models

Table 1 shows a summary of results on TIMIT in SV for three speaker models. The training speech of each model is changed from 2 to 20 seconds.

Table 1. EER for GMM-SVDD, GMM and SVDD with different training speeches in SV

| training speech | 2s | 4s | 6s | 8s | 10s |
|---|---|---|---|---|---|
| GMM-SVDD | 0.1192 | 0.0866 | 0.0672 | 0.0695 | 0.0658 |
| SVDD | 0.1304 | 0.1066 | 0.0953 | 0.1018 | 0.0848 |
| GMM | 0.1886 | 0.1277 | 0.1041 | 0.1019 | 0.1003 |
| training speech | 12s | 14s | 16s | 18s | 20s |
| GMM-SVDD | 0.0572 | 0.0486 | 0.0479 | 0.0468 | 0.0419 |
| SVDD | 0.0731 | 0.0662 | 0.0696 | 0.063 | 0.0571 |
| GMM | 0.0908 | 0.0855 | 0.0732 | 0.0754 | 0.0739 |

As listed in table 1, the performance of each model shows an upward trend with the number of training samples increased and GMM-SVDD overmatch other two models in all situations. SVDD outperform GMM in this experiment and achieves average improvements of up to 16% in EER over GMM. GMM-SVDD achieves average improvements of up to 37% in EER over GMM and 24% over SVDD in SV.

3.2 Speaker identification

In SI experiment, 50 speakers are selected and each one has 10 test speeches. The final system evaluation result derives from the error rate of 500 tests. Table 2 shows a summary of results on TIMIT in SI for three speaker models. The training speech of each model is changed from 2 to 20 seconds.

**Table 2.** Error rate for GMM-SVDD, GMM and SVDD with different training speeches in SI

| training speech | 2s | 4s | 6s | 8s | 10s |
|---|---|---|---|---|---|
| GMM-SVDD | 0.235 | 0.124 | 0.066 | 0.054 | 0.053 |
| SVDD | 0.274 | 0.212 | 0.202 | 0.204 | 0.18 |
| GMM | 0.498 | 0.228 | 0.112 | 0.11 | 0.088 |
| training speech | 12s | 14s | 16s | 18s | 20s |
| GMM-SVDD | 0.056 | 0.042 | 0.034 | 0.024 | 0.022 |
| SVDD | 0.174 | 0.176 | 0.169 | 0.148 | 0.134 |
| GMM | 0.066 | 0.056 | 0.038 | 0.034 | 0.028 |

As listed in table 2, the performance of each model shows an upward trend with the number of training samples increased and GMM-SVDD overmatch other two models in all situations. We can see that the SVDD results are significantly worse than the GMM results except in experiment of training with 2 seconds speeches. It is because that the hyper-spherical models of each SVDD are different, and after the introduction of fuzzy decision method, the score of each model are incomparable. Hence, further research about model normalization is need. GMM-SVDD achieves average improvements of up to 33% in error rate over GMM and 66% over SVDD in SI.

3.3 Speaker recognition

In SR experiment, 100 speakers are selected, 50 are selected as in-set speakers with the remaining 50 speakers taking on the role of out-of-set speaker (50in/50out). The final system evaluation result derives from the EER.

As can be seen from the above experiment, the error rate of GMM in SI is very low. In this section two fusion models are constructed, the first fusion model is mentioned above (fig.1), another is different from the former in identification phase. Flow chart of the second fusion model is shown in figure 4.

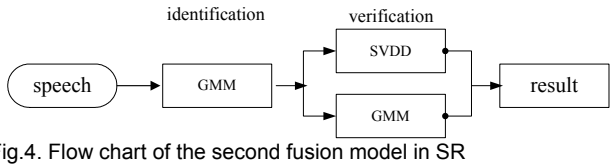


Fig.4. Flow chart of the second fusion model in SR

Table 3 shows a summary of results on TIMIT in SR for three speaker models. The training speech of each model is changed from 2 to 20 seconds.

As listed in table 3, the performance of each model shows an upward trend with the number of training samples increased and fusion models overmatch GMM in all situations. Fusion model 1 achieves average improvements

of up to 27% in EER over GMM and Fusion model 2 is up to 20% in SR. Introduction of SVDD in identification phase could bring average improvements of up to 9%.

Table 3. EER for fusion models and GMM with different training speeches in SR

| training speech | 2s | 4s | 6s | 8s | 10s |
|---|---|---|---|---|---|
| Fusion model 1 | 0.433 | 0.288 | 0.170 | 0.162 | 0.148 |
| Fusion model 2 | 0.480 | 0.331 | 0.204 | 0.18 | 0.162 |
| GMM | 0.562 | 0.374 | 0.244 | 0.215 | 0.201 |
| training speech | 12s | 14s | 16s | 18s | 20s |
| Fusion model 1 | 0.135 | 0.116 | 0.102 | 0.099 | 0.084 |
| Fusion model 2 | 0.142 | 0.125 | 0.104 | 0.106 | 0.092 |
| GMM | 0.173 | 0.159 | 0.149 | 0.134 | 0.128 |

## 4. Conclusions

In this paper, we presented a study of combination of two different models: GMM and SVDD. Our results indicate that the GMM-SVDD system fully utilizes the complementarity of SVDD and GMM and leads to significant gains over GMM or SVDD alone in SV, closed-set SI and SR.

Even though the SVDD clearly outperforms GMM in SV, however, in SI, SVDD results are significantly worse than the GMM. Next we consider the normalization of SVDD model to improve its performance in SI.


## Acknowledgment

The authors wish to acknowledge the financial support of Natural Science Foundation of Jiangsu Province in 2009(BK2009059)


## REFERENCES


[1] Reynold D. A., Quatieri T. F., Dunn R. B., “Speaker verification using adapted Gaussian mixture models”, Digital Signal Processing, 10(1-3),pp19-41, 2000.

[2] Tax D.M.J., Duin R.P.W., “Support vector data description”, Machine Learning, 54(1), pp45-66, 2004.

[3] Huenupán, F., Yoma N.B, Molina C. and Garreton C., “Confidence based multiple classifier fusion in speaker verification”, Pattern Recognition Letters, 29(7), pp957-966, 2008.

[4] L. Ferrer, Kemal Sönmez, and E. Shriberg, “An anticorrelation kernel for improved system combination in speaker verification”, In Proceedings of the Speaker and Language Recognition Workshop, Odyssey 2008, Stellenbosch, South Africa, January 2008b.

[5] Shou-Chun Yin, Rose, R., Kenny, P., “Adaptive score normalization for progressive model adaptation in text independent speaker verification”, IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, pp4857-4860, 2008.

[6] Banerjee A., Burlina P., Dieh C., “A support vector method for anomaly detection in hyperspectral imagery”, IEEE Transactions on Geoscience and Remote Sensing, 44(8), pp2282-91, 2006.

[7] Yuhuan Zhou, Xiongwei Zhang, Jinming Wang, Yong Gong, "Research on SVDD Applied in Speaker Verification", International Journal of Digital Content Technology and its Applications, 4(5), pp 89-95, 2010.

[8] Sturim D., Reynold D. A., “Speaker Adaptive Cohort Selection for T-norm in Text-Independent Speaker Verification”, IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, 2005.



***Authors**:*

*Yuhuan Zhou, Postgraduate Team 2, Institute of Communications Engineering, Biaoyin 2, Yudao Street, Nanjing, China, 210007, E-mail: zhouyh250@gmail.com;*

*Professor Xiongwei Zhang, Institute of Command Automation, PLA Univ. of Sci. & Tech..*

*Professor Jiming Wang, Institute of Communications Engineering, PLA Univ. of Sci.& Tech..*

*Yong Gong, Institute of Command Automation, PLA Univ. of Sci. & Tech..*

*Yi Zhou, Hydrometeorological Center of the South China Sea Navy.*