ON THE IMPACT OF USING OPTIMIZATION SEARCH METHOD IN LARGE SPEAKER DATABASE BASED ON HYBRID MODELING OVER EXPERIMENTAL INVESTIGATION

[1]Abdul Manan Ahmad and [2]Loh Mun Yee
Faculty of Computer Science and Information Systems
University Teknologi Malaysia
81300 Skudai, Johor

Email: [1]manan@utm.my, [2]lohmunyee@gmail.com

Abstract: Speaker recognition from speech signal is still an ongoing research in forensics and biometrics area. Speaker recognition is the process to enable machine to recognize speaker's identity from their speech. Recent development on classify speaker data from a group of speaker is still insufficient to provide a satisfied result in achieving high performance pattern classification engine. There are two main difficulties in this field: how to maintain accuracy rate under incremental amounts of training data and how to reduce the time processing in the case embedded systems need to consider about efficient and simplicity of calculation. Recently we have proposed three difference hybrid pattern classification approach for text independent speaker identification system; in these approaches, we combined a hybrid GMM/VQ and decision Tree model. The aim of this paper is to show the progress of the development of a high impact hybrid modeling. Besides, via this paper, an evaluation is done to verify the impact of using optimization search method on large speaker database.

Keywords: Speaker Identification System, Gaussian Mixture Model, Vector Quantization, Hybrid Vector Quantization/Gaussian Mixture Model

1. INTRODUCTION

The fast development of computing technology drives the expanding of database uses. The database systems have provided several of data for computing system; they are mushrooming and emerging in high speed to formulate the delivering, sharing and collaborating information. How to manage these large database or how to optimize the search process to retrieve a "right" data are current issue that always discussed by researchers.

A way that optimize the process of searching data by classify the data into smaller subgroup depends on it personal attribute have introduced via this paper. We believe that training data in small subgroup will towards the time efficiency if compare run the data searching under the whole set of data. This idea have implemented into speaker database which is under biometric security system, as known as speaker identification.

The evaluation of speech processing began after the development of channel vocoder also as know as voice coder and Voder (voice operated demonstrator). Until today, speech processing application are widely applying in many field like control automation system, biometric and forensic science authentication, human robot interacting and so on. Among these application, speech processing techniques are applying into security use. It formally knows as speaker recognition. Speaker recognition is a process where a person is recognized on the basis of his/her voice signals.

Speaker recognition can be text dependent or text independent. For text dependent system, ordinary a predefined utterance is used for training and for testing the system [1]. Whereas, in text independent, user can simply use whatever utterance they want for recognizes task.

Speaker verification and speaker identification are the subset of speaker recognition, where speaker verification accepts or rejects the identity claim of a speaker whereas speaker identification determines which registered speaker provides a given utterance from a set of know speaker [2]. The scope of our research is on the closed-set text-independent speaker identification task, which the close-set means the unknown voice must come from a fixed set of known speakers.

Speaker recognition is a quantum jump in artificial intelligence and forensic science technologies because it endows machines with the human-like abilities of distinguish people's identity from one another. To date, the recent technological and market growths of embedded systems draw our attention to enable speaker recognition systems running on embedded systems. Due to the memory and usage limitation of the embedded system, it cannot provide a complex calculation process for identify a speaker. Therefore, a design of a simple way for speaker classification techniques is needed. Recent development on classify speaker data from a group of speaker is still insufficient to provide a satisfied result in achieving high performance pattern classification engine. There are two main difficulties in this field: how to maintain accuracy rate under incremental amounts of training data and how to reduce the time processing in the case embedded systems need to consider about efficient and simplicity of calculation. These two criteria have to carry out to improve this research.

Dynamic Time Warping (DTW) [3], Vector Quantization (VQ) [4], Hidden Markov Models (HMM) [5], Gaussian mixture model (GMM) [6] and Support Vector Machine [7] are the most popular pattern classification techniques for speaker recognition. Other than these traditional methods, there are some hybrid methods as an alternative for speaker pattern classification. These hybrid method draw the attention of the researcher because it is proved that it bring a significant improvement for speaker recognition's research area. For example, they are the hybrid of GMM/ Neural network [8], hybrid of GMM/VQ [12, 13, and 14] and hybrid of GMM/SVM [9, 10, and 11]. Admittedly, the performance of speaker recognition systems in term of accuracy rates has been significantly improved over hybrid conditions. However, when speaker recognition is adopted in real-world application, time processing issue is often observed. Meanwhile, current works for the hybrid production of speaker recognition are almost directed towards accuracy problems, not time processing problems.

The aim of this paper is to show the progress of the development of a high impact hybrid modeling. Besides, via this paper, an evaluation is done to verify the impact of using optimization search method on large speaker database. Recently we have proposed three difference hybrid pattern classification approach for text independent speaker identification system; in these approaches, we combined a hybrid decision tree, GMM and VQ model. In this paper, we extend our investigations in order to select the most suitable hybrid model for real time application. In the work reported in this paper, we concern about a experimental investigation of three hybrid techniques for speaker identification. The emphasis of the experiments is on the accuracy of the models under incremental amounts of training data and the time taken to processing data.

This paper is organized as follows. In Section 2, reviews our proposed speaker identification framework and section 3, discusses about the research hypothesis and objective. Section 4 on the other hand, presents some basic of pattern classification techniques that using in the modeling and section 5 discusses our motivation of hybrid and how to construct three types of hybrid modeling for pattern classification. Section 6 shows the experimental result for these 3 techniques. Finally, section 7 is the conclusion.

2. SPEAKER IDENTIFICATION FRAMEWORK

The structure of the proposed speaker identification framework shows in figure 1. Speaker identification system involves two main stages, the enrollment stage and the identification stage. The speech signal is first processed to extract features that conveying speaker information. In the enrollment phase, all speaker data will be train by a pattern classification technique and save it into a bank of models. While in the identification stage, features are compared to a bank of models which are obtained from previous enrollment.

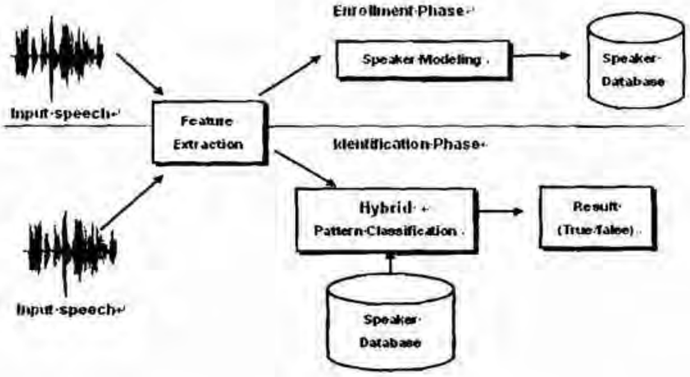

Figure 1. Structure of proposed speaker identification framework

3. RESEARCH HYPOTHESIS AND OBJECTIVE

Via this paper, a declaration has come out to justify that some preprocessing of classification process have to done on the large data set in order to different ship them via their attribute. By this classification process, data training and searching become faster because the engine just run on small range of data. This can do by hybrid method of pattern classification process. In the case of speaker identification system, speaker data can be divided into smaller subgroup followed by their gender, which is the attribute of the speaker data. Our objective over the research is classification the speaker data using hybrid method. By using hybrid method, we aims to process huge speaker data in short time limit and in the same time we manage to gain better accuracy rate for speaker data searching.

4 THE BASIC OF PATTERN CLASSIFICATION TECHNIQUES

4.1 Gaussian Mixture Models (GMM)

The Gaussian mixture model (GMM) is a density estimator and is one of the most commonly used types of classifier. In this method, the distribution of the feature vector x is modeled clearly using a mixture of M Gaussians.

$$P(x|M)=\sum_{i=1}^{M} u_i \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \tag{1}$$

Here mu_i, Σ_i represent the mean an d covariance of the i^{th} mixture. Given the training data x_1, $x_2 \ldots x_n$, and the number of mixture M, the parameters μ_i, Σ_i, α_i is learn using expectation maximization. During recognition, the input speech is again used extract a sequence of features x_1, $x_2 \ldots x_L$. the distance of the given sequence from the model is obtained by computing the log likehood of given sequence given the data. The model that provies most highest likelihood score will verify as the identity of the speaker. A detailed discussion on applying GMM to speaker modeling can be found in [15].

4.2 Vector Quantization (VQ)

Vector Quantization (VQ) is a pattern classification technique applied to speech data to form a representative set of features. Among the first apply this technique to speaker verification were Soong et al (1985) and Buck et al (1985) [20, 21].

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook.

In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result code words (centroids) are shown in Figure 2 by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with the smallest distortion is identified.

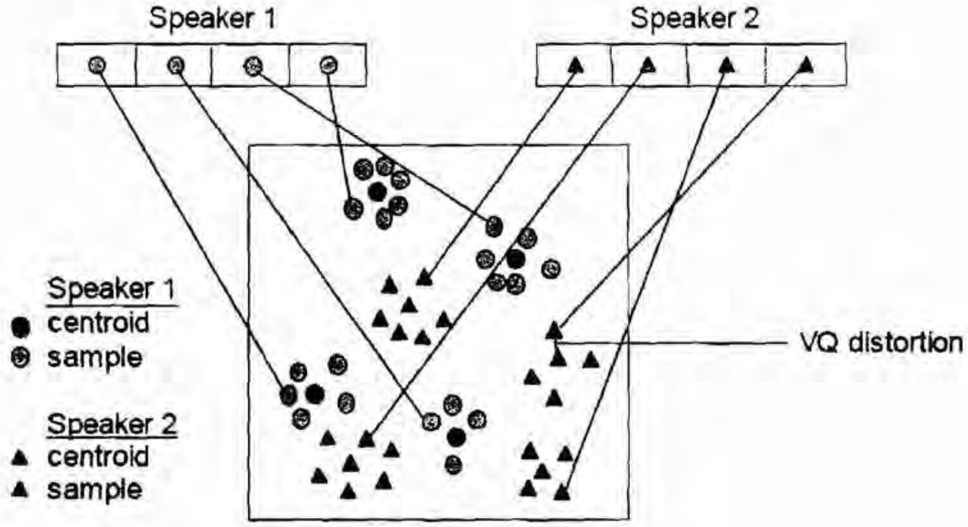

Figure 2 Structure of how VQ work as classifier

5 A MOTIVATION OF HYBRID CLASSFIER

This research aims to maintain accuracy under incremental amounts of training data and reduce the processing time for real time systems which need to consider about efficiency and simplicity of calculation. Most of the speaker recognition system use GMM as pattern classifier. This technique introduce by Reynolds [15] since the year of 1995. GMM method represents each speaker data into different GMM to generate a speaker model for training and testing. It calculates the log likelihood score for all training speaker data and makes the decision followed by maximum posteriori probability. Even though it gain high accuracy, but the calculation for all training data make it become complexity.

Due to the above reason, we have try several hybrid modeling to solve the problem. From the research, we introduce the decision tree modeling to scope the speaker data for training purpose in order to separate it into smaller group. Besides, we also try apply the VQ technique into the hybrid modeling since VQ are proved more simplicity calculation then GMM because it just depends on the cookbook size [16]. There are three ways of hybrid have explore; whereas the aims of this paper is to investigate the most suitable framework for real time application.

5.1 Parallel hybrid VQ/GMM classifier

In the first proposed hybrid modeling, hereinafter referred to as "VQ+GMM 1", both VQ model and GMM model will run parallel after signal preprocessing process [17]. Figure 3 shows an overview process of this VQ+GMM 1 hybrid modeling.

In GMM training phase, an MFCC output will return as GMM input after compute signal Mel-frequency cepstrum coefficients. For speaker identification, each speaker is represented by a GMM and is referred to by his/her speaker model. GMM classification engine will calculate log likelihood score for all training speaker data and save it into a speaker model. While in testing phase, a comparison about training speaker and testing speaker will be done. GMM classification engine will make a decision followed by maximum posteriori probability.

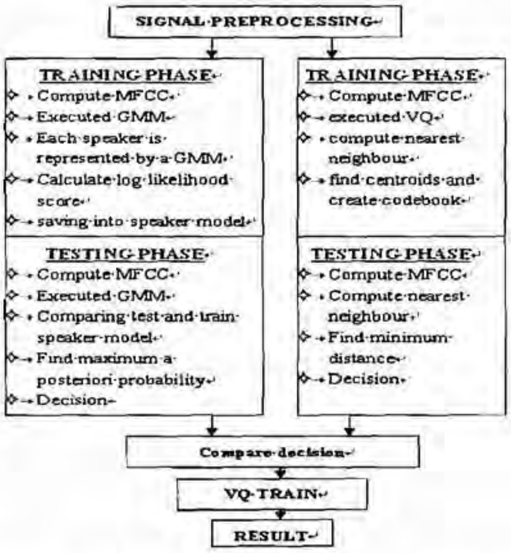
SIGNAL PREPROCESSING
TRAINING PHASE
Compute MFCC
Executed GMM
Each speaker is represented by a GMM
Calculate log likelihood score
saving into speaker model
TESTING PHASE
Compute MFCC
Executed GMM
Comparing test and train speaker model
Find maximum a posteriori probability
Decision
TRAINING PHASE
Compute MFCC
executed VQ
compute nearest neighbour
find centroids and create codebook
TESTING PHASE
Compute MFCC
Compute nearest neighbour
Find minimum distance
Decision
Compare decision
VQ TRAIN
RESULT

Figure 3. First hybrid method, VQ+GMM 1, speaker identification system based on parallel VQ/GMM classifier

In VQ training phase, Vector Quantization using Linde-Buzo-Gray algorithm is executed using MFCC as input. Later on, we will run the nearest-neighbour search to find the codeword in the current codebook that is closest and assign that vector to the corresponding cell. Then, the system will find centroids and update for each speech signal and the cookbooks are created. In testing phase, a function will computes the Euclidean distance between training data and testing data. The system will identify which calculation yields the lowest value and checks this value against a constraint threshold. If the value is lower than the threshold, the system outputs an answer. After the GMM and VQ engine come out a decision, both result will be compare in decision logic. If there are divergences between them, both speaker data will be train and test again in a VQ classification machine. We have to mention that, in this phase, our identification scopes already deflate to two speakers, therefore it will gain a high accuracy rate. Besides, the main reason we choose VQ for second phase classifier was on account of it offers simplicity in computation.

5.2 Vector Quantization Decision Tree Modeling

The second hybrid method was using decision tree (DT) theory and VQ approach, hereinafter referred to as "DT+VQ 2" [18]. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision.

Decision trees represent rules. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved. In our proposed modeling, we take the superiority of decision tree theory, which is simplicity computation to distinguish a group of speaker into smaller subgroup. We believe that smaller group of training data will decrease time processing in run times.

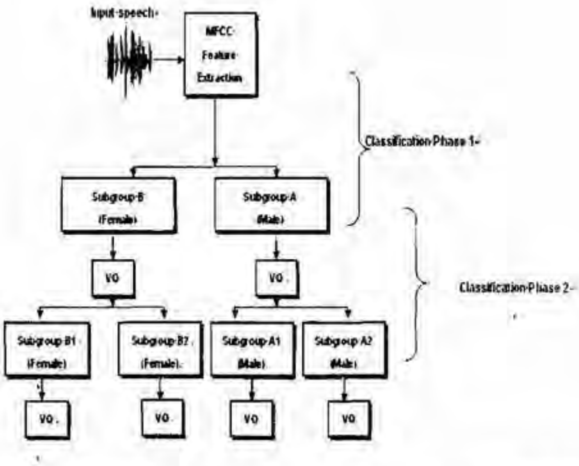

Figure 4. Second hybrid method, DT+VQ 2, speaker identification system based on decision tree theory and VQ approach

The overall structure of our hybrid system is depicted in figure 4. After MFCC feature extraction process, the speech signal will transform to a feature vector form. For the phase 1 of the classification, we use decision tree theory to classify out the gender of the speaker in order to group them into smaller group. Once a decision tree has been built, it is used as a component of the complete classification system. Normally identification errors for huge database often occur when a speaker is taken for another speaker belonging to the same gender. For example male speaker A unrecognized as another male speaker B.

In phase 2 classification, we use the decision tree function to separate out the speaker model that gain the similar score into 4 difference group which are subgroup A1, A2, B1 and B2. This process aims to solve the similarity speaker problem in order to make an improvement on the accuracy rate when the application facing a huge database. Later on, the VQ classification progress will only done in one particular subgroup which the group has identify as it store the speaker model. The system will identify which calculation yields the lowest value and checks this value against a constraint threshold. If the value is lower than the threshold, the system outputs an answer.

5.3 Vector Quantization Decision Rules for Gaussian Mixture Modeling

In the third type of proposed hybrid modeling, hereinafter referred to as "VQ/DT+GMM 3", the novel model is an extended version for DT+VQ 2, which add on one more phase classification using GMM approach [19]. The reason of using GMM as classifier in the subgroup because of after the experiments of DT+VQ 2, we found that some error of classification that cannot recognized by VQ techniques still remain. Therefore, the third phase classification, we apply the GMM technique.

The overall structure of our hybrid system is depicted in figure 5. After separate the speaker model into 4 small subgroup using VQ/DT as decision function, we apply GMM classification for finding the identity of the speaker. On account of the GMM model just need to train speaker data in the subgroup instead training all speaker data, the computation time will decrease.

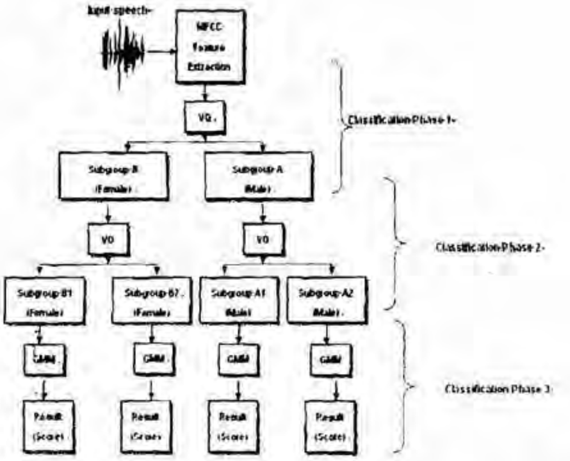

Figure 5. Third hybrid method, VQ/DT+GMM 3, speaker identification system based on VQ decision rules for Gaussian mixture modeling

6. EXPERIMENTAL SETUP AND RESULT

In orders to get a fair comparison between 3 types of classifier, experiments are conducted on a clean condition. We performed our evaluation on the TIMIT speech database. Out of this large set, we chose 5 utterances of 10-100 distinct users to evaluate our system. The emphasis of the experiments is on the accuracy of the models under incremental amounts of training data and the time taken to processing data.

6.1 System evaluation based on accuracy over incremental data

The first method evaluated uses VQ+GMM 1 as pattern classification techniques. For consistent reason, the experiments have run 5 times and the averages result obtained. Table 1 shows the effect of increasing the speakers on performance of the VQ+GMM 1 speaker identification system. Accuracy starts off highly 100% as would be expected, and slowly declines to approximately 96% when testing data increased. As can be observed, VQ+GMM 1 speaker verification accuracy rate has decrease when the training data increase; this is due to the complexity of the computation. Through the experiments, the VQ+GMM 1 obtain 97.44% for accuracy rates over an average amount.

The second method evaluated uses DT+VQ 2 as pattern classification techniques. Table 2 shows the effect of increasing the speakers on performance of the DT+VQ 2 speaker identification system for speakers from 10 to 100. Accuracy starts off highly 100%, and slowly declines to approximately 97%. As can be observed, DT+VQ 2 obtain the better result if compare with VQ+GMM 1, that is 98.26% for accuracy rates over an average amount.

The third method evaluated uses VQ/DT+GMM 3 as pattern classification techniques. Table 3 shows the effect of increasing the speakers on performance of the VQ/DT+GMM 3 speaker identification system for speakers from 10 to 100. Accuracy starts off highly 100%, and slowly declines to approximately 98%. As can be observed, even VQ/DT+GMM 2 speaker identification accuracy rate has decrease when the training data increase, but it still obtain the better result if compare with DT+VQ 2, that is 98.68% for accuracy rates over an average amount..

Table 1. Performance of the VQ+GMM 1 on increasing testing data

No. of testing data	Average no. of false identification	Accuracy Percentages (%)
10	0	100%
20	0	100%
30	0	96%
40	1	97.5%
50	2	96%
60	2	96.67%
70	2	97.14%
80	2	97.5%
90	3	96.67%
100	3	97%

Table 2. Performance of the DT+VQ 2 on increasing testing data

No. of testing data	Average no. of false identification	Accuracy Percentages (%)
10	0	100%
20	0	100%
30	0	100%
40	1	97.5%
50	1	96%
60	1	98.3%
70	1	98.57%
80	2	97.5%
90	2	97.78%
100	3	97%

Table 3. Performance of the VQ/DT+GMM 3 on increasing testing data

No. of testing data	Average no. of false identification	Accuracy Percentages (%)
10	0	100%
20	0	100%
30	0	100%
40	0	100%
50	1	98%
60	1	98.33%
70	2	97.14%
80	2	97.5%
90	2	97.78%
100	2	98%

6.2 System evaluation based on time complexity analysis

The emphasis of the experiments for this section is on the time taken to processing data between both VQ+GMM 1 classifier, DT+VQ 2 classifier and VQ/DT+GMM 3 classifier. The result of time processing for 10 speakers by using each techniques shows in table 4. We report that the VQ+GMM 1 need 78.23 seconds for the whole training and testing process whereas DT+VQ 2 and VQ/DT+GMM just need 41.76 seconds and 50.42 seconds. Thus, the third method, VQ/DT+GMM 3 are more suitable to apply in real time because of it accuracy and processing time. Obviously, a significant improvement compared to the VQ/DT+GMM 3 is reported, a reduction in identification times up to 25% is reached.

Table 4. VQ+GMM 1 and VQ+GMM 2 comparative result for time processing

Algorithm	*VQ+GMM 1*	*DT+VQ 2*	*VQ/DT+GMM 3*
Time	78.23sec	41.76sec	50.42sec
Number of speaker	10	10	10

6.3 Discussion

The observation has done by the first experiment. From there, the system takes the longest time to process because the two pattern classifier run parallel. From experiment, we notice that the searching process make the system slow down. Therefore, an optimize search method are adding in the second modeling and we done some arrangement on the modeling. However, there are some weaknesses or some identification errors occur when doing the experiment for the second method. The main reason is even VQ are proven simplicity for calculation, but the ability for identification are poor. Consequently, the third model is designed by adding GMM as classification function.

7. CONCLUSIONS

In this paper, an investigation over 3 types of hybrid modeling, VQ+GMM 1, DT+VQ 2 and VQ/DT+GMM 3 has done. This investigation intend to make a comparison among these hybrid modeling in order to decide he most suitable hybrid model for real time application. From the experiments, we observe that one good way of applying hybrid method between VQ and GMM because of their difference ways to classified data.

The result of the experiments have shown that the VQ/DT+GMM 3 classifier always yields better result if compare to other hybrid classifier. Perhaps the most surprising overall finding presented in this paper is the superior performance of time processing over VQ/DT+GMM 3 classifier. Therefore, this classifier should be considered for real world application.

Future work will be concentrating on investigation of the effectiveness of feature extraction techniques for more robust speaker recognition. Investigation on a better adaptation function also will be done to ensure that the hybrid classifier get the better accuracy rate.

ACKNOWLEDGEMENTS

This work was jointly supported by IRPA e-Science Fund 79087. The authors would like to thank the Ministry of Science; Technology & Innovation (MOSTI) of Malaysia for without the sponsorships, the authors could not have materialized our ideas and visions. In addition, the authors are grateful to University Technology of Malaysia (UTM Skudai, Johor) for providing all the facilities and equipments throughout the entire course and the project.

REFERENCES

[1] Reynolds, D.A., "An overview of automatic speaker recognition technology", Proc. of the ICASSP '02, IEEE International Conference on Acoustics, Speech, and Signal Processing Volume: 4, 2002, pages: IV-4072- IV-4075.

[2] Campbell, J.P., "Speaker Recognition: A Tutorial", Proc. of the IEEE, vol. 85,no. 9, 1997, pages. 1437-1462.

[3] Sakoe, H.and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", Acoustics, Speech, and Signal Processing, IEEE Transactions on Volume 26, Issue 1, Feb 1978, Page 43 - 49.

[4] Vlasta Radová and Zdenek Svenda, "Speaker Identification Based on Vector Quantization", Proceedings of the Second International Workshop on Text, Speech and Dialogue, Vol. 1692,1999, Pages: 341 - 344.

[5] Lawrence R. Rabiner., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, 77 (2), 1989, p. 257–286.

[6] Reynolds, D. A. and Rose, R. C. "Robust text-independent speaker identification using Gaussian mixture speaker models". IEEE Trans. Speech Audio Process. 3, 1995, pp 72–83.

[7] Solera, U.R., Martín-Iglesias, D., Gallardo-Antolín, A., Peláez-Moreno, C. and Díaz-de-María, F, "Robust ASR using Support Vector Machines", Speech Communication, Volume 49 Issue 4, 2007.

[8] M. F. BenZeghiba and H. Bourlard. "User-Customized Password Speaker Verification based an HMM/ANN and GMM models". Proceedings of ICSLP 2002, pp 1325-1328.

[9] S. Fine, J. Navratil and R. A. Gopinath, "Enhancing GMM scores using SVM "hints"," in Proceedings of Eurospeech, 2001, pp. 1757-1761.

[10] Minghui Liu, Yanlu Xie, Zhiqiang Yao, Beiqian Dai, "A New Hybrid GMM/SVM for Speaker Verification", Proc. of the ICPR 2006, 18th International Conference on Pattern Recognition Volume: 4, 2006, pages: 314-317.

[11] Fenglei Hou and Bingxi Wang, "Text-independent speaker recognition using probabilistic SVM with GMM adjustment", Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, 26-29 Oct. 2003, Pages: 305 - 308

[12] J. Pelecanos, S. Myers, S. Sridharan and V. Chandran, "Vector Quantization Based Gaussian Modeling for Speaker Verification", 15th International Conference on Pattern Recognition, Volume 3, 2000,p. 3298.

[13] Qiguang Lin, Ea-Ee Jan, ChiWei Che, Dong-Suk Yuk and Flanagan, J, "Selective use of the speech spectrum and a VQGMM method for speaker identification", Fourth International Conference on Spoken Language, Vol 4, 1996, Pg:2415 - 2418.

[14] Singh, G.; Panda, A.; Bhattacharyya, S.and Srikanthan, T. , "Vector quantization techniques for GMM based speaker verification", IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 2, Issue , 6-10 April 2003, Page(s): II - 65-8.

[15] Reynolds, D. A. and Rose, R. C. "Robust text-independent speaker identification using Gaussian mixture speaker models",IEEE Trans. Speech and Audio Process. 3, 1995, pp 72–83.

[16] Yu, K., Mason, J., Oglesby, J., "Speaker recognition using hidden Markov models, dynamic time warping and vector quantization", Vision, Image and Signal Processing, IEE Proceedings, Oct 1995.

[17] Loh Mun Yee and Abdul Manan Ahmad, "Text-Independent Speaker Identification Using Hybrid Vector Quantization / Gaussian Mixture Models Pattern Classifier", Proceedings of International Conference on Control, Automation and Systems 2008, Oct. 14-17,2008.

[18] Loh Mun Yee and Abdul Manan Ahmad, "Towards Making Better Decision Rule For Speaker Identification Via Decision Tree Theory ", ICSTIE2008, International Conference on Science and Technology 2008, 12-13 of December 2008, UiTM Penang, Malaysia, in press..

[19] Loh Mun Yee, Abdul Manan Ahmad, "Vector Quantization Decision Function for Gaussian Mixture Model Based Speaker Identification", ISPACS 2008, Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems, December 8-11, 2008, Bangkok, Thailand , in press.

[20] Soong, F. K., Rosenberg, A. E., Rabiner, L. R. and Juang, B. H. (1985). "A vector quantization approach to speaker recognition." Proc. ICASSP'85, 387-390.

[21] Buck J. T., Burton D. K. and Shore J. E. (1985). "Text-dependent speaker recognition using vector quantization." Proceedings of ICASSP-85, 1:391-394.