COMPREHENSIVE DESIGN AND DEVELOPMENT OF TIME EFFICIENCY SPEAKER RECOGNITION MODEL FROM FRONT END TO BACK END

[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: The rapid development of the forensic science technologies has been evolved speaker recognition to becoming one of the research topics. However, pattern classification from speech signal remains as challenging problem encountered in general speaker recognition system, including speaker verification and speaker identification. Conventional speaker recognition researches are almost directed towards accuracy problems, not time processing problems. Due to the needs of reduction time processing of speaker recognition system, this research focuses on develop a comprehensive design of speaker recognition model from front end to back end which able to process speaker data in short time limit. In the front end process, we introduce some pre-processing techniques to enhance the speech signal. Whereas, for the back end process, we propose a decision function by using vector quantization techniques to decrease the training model for GMM in order to reduce the processing time. Experimental result shows that our hybrid VQ/GMM method always yielded better improvements in accuracy and bring almost 30% reduce in time processing. In this paper, a new, robust and simplicity computation method of pattern classification technique for speaker identification system is proposed. Consequently, this research is intended to develop a fully optimize ways speaker identification approach from hybrid modeling.

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

1. INTRODUCTION

Speaker recognition is a process where a person is recognized on the basis of his/her voice signals [1]. Speaker recognition can be further broken into two categories: speaker identification and speaker verification. Identification takes the speech signal from an unknown speaker and compares this with a set of valid users. The best match is then used to identify the unknown speaker. Similarly, in verification the unknown speaker first claims identity, and the claimed model is then used for identification. If the match is above a predefined threshold, the identity is accepted. A complete speaker recognition system consists of front end and back end process. The front end which contain preprocessing and feature extraction while the back end which contain pattern classification techniques.

Feature extraction plays as a crucial part in speaker recognition component chain. The goal of this stage is to extract speaker dependent information from speech signal and represent it by a set of vectors called feature. Since this stage is the first component in the chain, the quality of this stage will strongly affect the quality of other components (speaker modeling and pattern matching).

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern classification. The goal of pattern classification is to classify objects of interest into a number of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech. The classes here refer to individual speakers [2]. Pattern classification plays as a crucial part in speaker modeling component chain. The result of pattern classification will strongly affect the speaker recognition engine to decide whether to accept or reject a speaker.

Many research efforts have been done in speaker recognition pattern classification. There are Dynamic Time Warping (DTW) [3], Vector Quantization (VQ) [4], Hidden Markov Models (HMM) [5], Gaussian mixture model (GMM) [6] and so forth. There are some weaknesses in these techniques. Consequently, Support Vector Machine (SVM) was introducing as an alternative classifier for speaker verification [7]. SVM, which are based on the principle of structural risk minimization, consist of binary classifiers that maximize the margin between two classes. The power of SVM lies in their ability to transform data to a higher dimensional space and to construct a linear binary classifier in this space. It sounds efficient and useful to speaker recognition application, but they cannot easily deal with the dynamic time structure of sounds, since they are constrained to work with fixed-length vectors [8]. When working with audio signals, each signal frame is converted into a feature vector of a given size, the whole acoustic event is represented by a sequence of feature vectors, which shows variable length. However, SVM which only work with fixed-length vectors means that it only can accept text dependent for training and testing data.

Despite the extensive research has been performed in speaker recognition area over the last few years, it still remain a great challenge in managing and process huge speaker data sets in a short time limit. Conventional speaker recognition researches are almost directed towards accuracy problems, not time processing problems. Due to the needs of reduction time processing of speaker recognition system, this research focuses on develop a comprehensive design of speaker recognition model from front end to back end which able to process speaker data in short time limit. In the front end process, we introduce some pre-processing techniques to enhance the speech signal. Whereas, for the back end process, we propose a decision function by using vector quantization techniques to decrease the training model for GMM in order to reduce the processing time.

In the work reported in this paper, we also concerns on comparison of DTW, VQ, GMM, SVM and our hybrid pattern classifier for speaker recognition. The emphasis of the experiments is on the performance of the models under incremental amounts of training data in an attempt to identify the best approach for speaker recognition in order to improve the problem as just stated as paragraph above.

This paper is organized as follows. In Section 2, reviews the propose speaker recognition structure. In Section 3, discusses the methods which use for preprocessing signal and section 4 shows the feature extraction techniques. Section 5 discusses how we construct DTW, VQ, GMM, SVM and our hybrid method for speaker recognition. Section 6 shows the experimental result for these 5 techniques. Finally, section 7 concludes our work.

2. OUR SPEAKER RECOGNITION FRAMEWORK

In this section, we generally reviews our propose speaker recognition structure. Our Speaker recognition system involves two main stages, the enrolment stage and the verification stage. These phases involve three main parts:

- Pre-Processing.
- Feature Extraction.
- Pattern Classification.

A block diagram of this procedure is shown in Figure 1. At the time of enrollment, speech sample is acquired in a controlled and supervised manner from the user. The speaker recognition system has to process the speech signal in order to extract speaker discriminatory information from it. This discriminatory information will form the speaker model. At the time of verification a speech sample is acquired from the user. The recognition system has to extract the features from this sample and compare it against the models already stored before hand. This is a pattern matching or classification task.

Feature extraction maps each interval of speech to a multidimensional feature space. This sequence of feature vectors X_i is then compared to speaker models by pattern classification. This results in a match score Z_i for each vector or sequence of vectors. The match score measures the similarity of the computed input feature vectors to models of the claimed speaker or feature vector patterns for the claimed speaker. Last, a decision is made to either accept or reject the claimant according to the match score or sequence of match scores, which is a hypothesis testing problem.

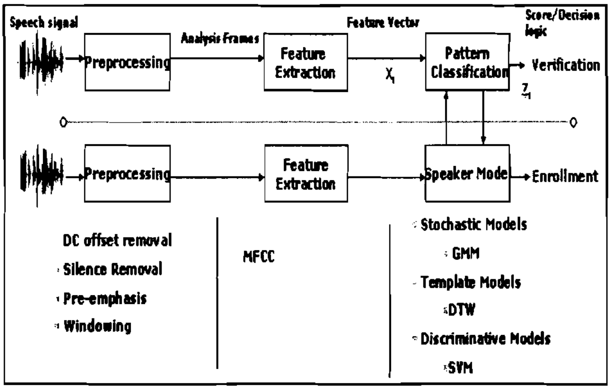

Figure 1. Speaker Recognition Framework

3. PRE-PROCESSING

All speech data will perform in a discrete-time speech signal because of recorded by sampling the input. Therefore, we need some pre-processing techniques to make the discrete-time speech signal more flexible for the processes that follow. There are 4 pre-processing techniques that we before feature extraction. These include DC offset removal, silence removal, pre-emphasis and windowing.

3.1 DC Offset Removal

Speech data are discrete-time speech signal, it often carry some redundant constant offset called DC offset [9]. These DC offset will effect quality of the information extracted from the speech signal. Consequently, we calculating the average value of the speech signal and subtracting this from itself.

3.2 Silence Removal

This process is performed to discard silence periods from the speech containing silence frames. So, the signal becomes more compact as shown in the Figure 3. Silence frames are audio frames of background noise with a low energy level with respect to voice segments. The signal energy in each speech frame is evaluated by equation (1).

$$E_i = \sqrt{\sum_{k=1}^{M} x_i(k)^2} \qquad i = 1,\ldots\ldots,N \tag{1}$$

Where M is the number of samples in a speech frame and N is the total number of speech frames. Threshold is successively performed to detect silence frames with a threshold level determined by equation (2).

$$\text{Threshold} = E_{min} + 0.1\ (E_{max} - E_{min}) \tag{2}$$

$E_{max} - E_{min}$ are the lowest and greatest values of the N segment respectively.

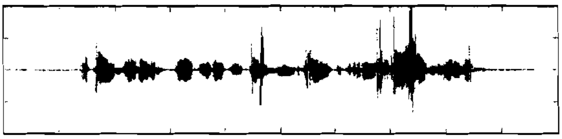

Figure 2. Speech Signal before Silence Removal

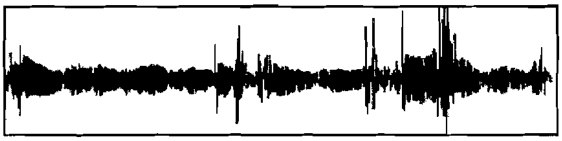

Figure 3. Speech Signal after Silence Removal

3.3 Pre-emphasizing

Pre-emphasis is a technique used in speech processing to enhance high frequencies of the signal. The main purpose of pre-emphasizing is to spectrally flatten the speech signal that is to increase the relative energy of its high-frequency spectrum.

There are two important factors driving the need for pre-emphasis. Firstly, the speech signal generally contains more speaker specific information in the higher frequencies [10]. Secondly, as the speech signal energy decreases the frequency increases. This allows the feature extraction process to focus on all aspects of the speech signal. Pre-emphasis is implemented as a first-order Finite Impulse Response (FIR) filter defined as:

$$H(Z) = 1\text{-}0.95\ Z^{-1} \tag{3}$$

Figure 4 shows the speech signal before pre-emphasizing process and Figure 5 shows the speech signal after pre-emphasizing process

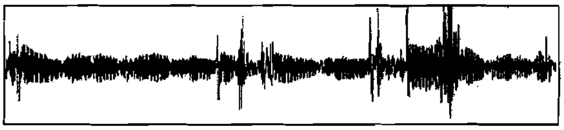

Figure 4. Speech Signal before Pre-emphasizing

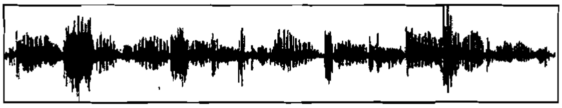

Figure 5. Speech Signal after Pre-emphasizing

3.4 Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. A windowing function is used on each frame to smooth the signal and make it more amendable for spectral analysis. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as, where N is the number of samples in each frame, then the result of windowing is the signal.

$$y_l(n) = x\,(n)w(n), \qquad 0 \leq n \leq N\text{-}1 \tag{4}$$

Typically the Hamming Window is used, which is of the form

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N\text{-}1 \tag{5}$$

4.0 FEATURE EXTRACTION

After Davis and Mermelstein reported that Mel-frequency cepstral coefficients (MFCC) provided better performance than other features in 1980 [11], MFCC has been widely used as the feature parameter for automatic speaker recognition. In our implementation, we will use MFCC technique[21] to extract the speech feature in order to obtain the best result for pattern classification. Figure 6 shows an outline of the process of MFCC.

MFCC start with dividing the speech signal into short frame and windowing each frame to discard the effect of discontinuities at edges of the frames. In fast fourier transform (FFT) phase, it convert the signal to frequency domain and after that Mel scale filter bank is applied to the resulting frames. After Mel frequency warping the frames, logarithm of the signal is passed to the inverse DFT function converting the signal back to time domain. As a result of the final step, 13 coefficients named MFCC for each frame are obtained. The 0th coefficient is not used because it represents the average energy in the signal frame and contains little or no usable information.

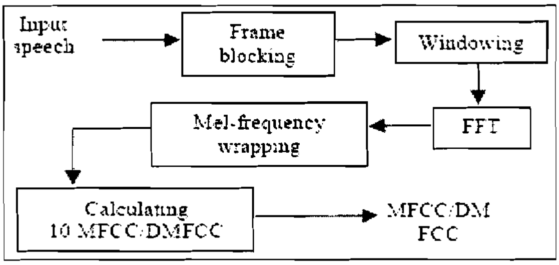

Figure 6. Outline of the process of MFCC

As the output of feature extraction phase, vectors in 12 dimensions are obtained for each frame. These vectors are used in pattern matching/classification technique for compare and match the feature sets against the model already stored before hand.

5.0 PATTERN CLASSIFICATION

The pattern classification task of speaker recognition involves computing a match score, which is a measure of the similarity of the input feature vectors to some model. Speaker models are constructed from the features extracted from the speech signal. To enroll users into the system, a model of the voice, based on the extracted features, is generated and stored (possibly on an encrypted smart card). Then, to authenticate a user, the matching algorithm compares/scores the incoming speech signal with the model of the claimed user.

In our experiment, four major speaker recognition pattern classification techniques has been chosen and a comparison of the performance with proposed technique has made. These techniques are DTW, GMM, VQ and SVM.

5.1 Four Major Speaker Recognition Pattern Classification Techniques

5.1.1 Dynamic Time Warping

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. In our experiment, we use the DTW techniques which propose by Sadaoki Furui at years 1981[12]. According to Furui teory, the training data are used as a initial template, and the testing data is time aligned by DTW. DTW is a method that allows a computer to find an optimal match between two given sequences. The average of the two patterns is then taken to produce a new template to which a third utterance is time aligned.

This process is repeated until all the training utterances have been combined into a single template.

The idea of the DTW technique is to match a test input represented by a multi-dimensional feature vector T= [t_1, t_2...t_I] with a reference template R= [r_1, r_2...r_j]. While aim of DTW is to find the function w(i). as shown in figure 7.

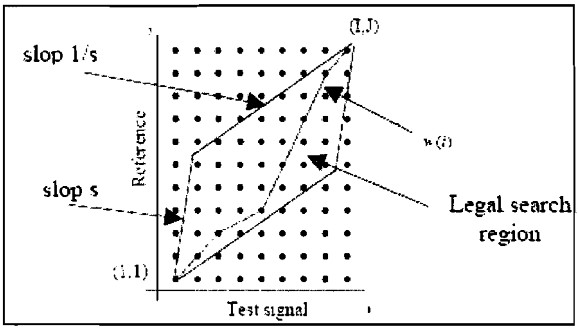

Figure 7. Idea of DTW

5.1.2 Gaussian Mixture Models

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.

$$P(x|M) = \sum_{i=1}^{m} a_i \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \qquad (6)$$

Here mu_i, Σ_i represent the mean an d covariance of the i^{th} mixture. Given the training data x_1, x_2...x_n, and the number of mixture M, the parameters μ_i, Σ_i, a_i is learn using expectation maximization. During recognition, the input speech is again used extract a sequence of features x_1, x_2...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 [6].

Figure 8 shows a process flow for GMM approach in speaker identification training phase and testing phase. In GMM training phase, an MFCC output will return as GMM input after compute signal Mel-frequency cestrum 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. The model that provides highest likelihood score will verify as the identity of the speaker.

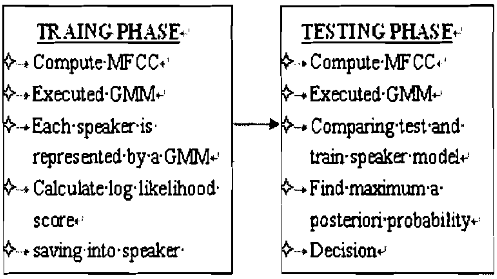
TRAING PHASE
Compute MFCC
Executed GMM
Each speaker is represented by a GMM
Calculate log likelihood score
saving into speaker
TESTING PHASE
Compute MFCC
Executed GMM
Comparing test and train speaker model
Find maximum a posteriori probability
Decision

Figure 8. A process flow of GMM in training and testing phase for speaker identification system.

5.1.3 Vector Quantization

Vector Quantization (VQ) is a pattern classification technique applied to speech data to form a representative set of features. It maps vectors to smaller regions called cluster. These cluster's center, centroid, are collected and will make up a codebook. The VQ codebook will represents the speaker feature from the training data. The speaker identification engine are depends on the codebook to identify a speaker. Figure 9 shows the speaker identification process flow for VQ in training and testing phase.

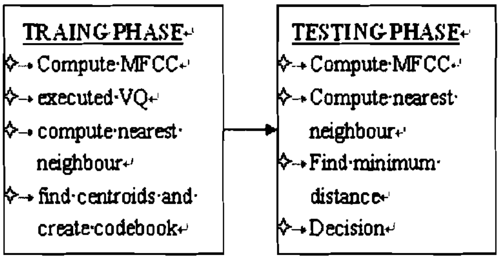

Figure 9. A process flow of VQ in training and testing phase for speaker identification system.

In VQ training phase, Vector Quantization is executed using MFCC as input. Later on, the speaker identification engine 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, its find centroids and update for each speech signal and the codebooks 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.

5.1.4 Support Vector Machines

SVM is a binary classification method that finds the optimal linear decision surface based on the concept of structural risk minimization. The decision surface is a weighted combination of elements of a training set. These elements are called support vectors, which characterize the boundary between the two classes. Let the two classes of the binary problem be labeled +1 and -1.

For the purpose to characterize the boundary between the two classes, we need maximizing the margin. Maximizing the margin are the process find the "middle-line" consider two parallel lines both of which separate the two classes without error. Several steps need to be determine the linear separator (Figure 10a, 10b, 10c) :

- Find closest points in convex hulls
- Plane bisect closest points
- Maximize distance between two parallel supporting planes

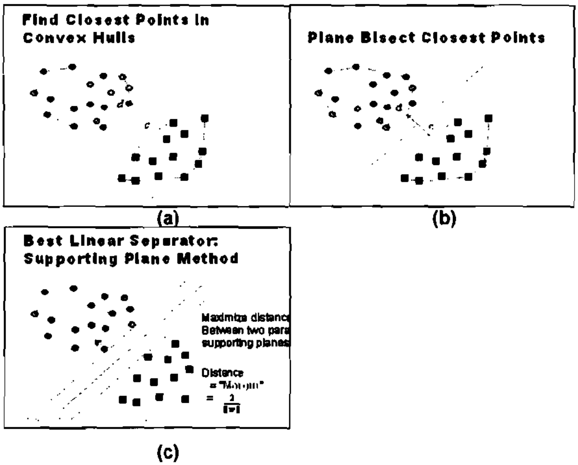
Find Closest Points In Convex Hulls
c
d
(a)
Plane Bisect Closest Points
c
d
(b)
Best Linear Separator: Supporting Plane Method
Maximize distanc
Between two para
supporting planes
Distance
= "Margin"
= $\frac{2}{\|w\|}$
(c)

Figure 10. Steps for Binary Linear Decision Boundary

During speaker recognition process, classifying the feature which derived from the transformation of feature extraction directly will not immediately works when using SVM [13]. It is because SVM only can process fixed-length input, whereas speech signals are non-stationary. Therefore, we need to categorizes the feature and scaling them.

SVM requires that each data instance is represented as a vector of real numbers. Hence, if there are categorical attributes, we first have to convert them into numeric data. We recommend using m numbers to represent an m-category attribute. Only one of the m numbers is one, and others are zero. For example, a two-category attribute such as {speaker, imposter} can be represented as (0,1) and (1,0).

Scaling them before applying SVM is very important. The main advantage is to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Because kernel values usually depend on the inner products of feature vectors, e.g. the linear kernel and the polynomial kernel, large attribute values might cause numerical problems. We recommend linearly scaling each attribute to the range [−1, +1] or [0, 1].

5.2 Motivation of Hybrid Model

The result of pattern classification will strongly affect the speaker recognition engine to decide whether to accept or reject a speaker. Early pattern classification is produced by Dynamic Time Warping and Hidden Markov Models. These techniques are not really efficient for real time application due to characteristic of text dependent recognition. As an alternative to solve text dependent problem, Vector Quantization (VQ), Gaussian mixture model (GMM) and Support Vector Machine was introduced for speaker recognition. GMM was a focus of research after Douglas Reynolds proves its effective performed in text independent speaker identification [6].

Besides, GMM are base on probabilistic framework, it provide high-accuracy recognition. For speaker identification task using GMM approach, each speaker data is modeled by a GMM; during testing phase, each GMM is calculated independently to estimate the parameters and compare with other GMM to find the best match score.

Although the GMM technique of pattern classification appears to have many advantages, however, in practice the process does not always produce satisfied result due to the long time processing. Consequently, alternative methods must be sought in order to reduce time processing problem. In these circumstances, pattern classification engine for speaker recognition should capable to manage and process huge speaker data sets in a short time limit. Meanwhile, current works for the production of speaker recognition are almost directed towards accuracy problems, not time processing problems. Therefore, it is encouraging if a speaker recognition task can be conducted in a "good" pattern classification machine.

In this paper, we propose a decision function by using vector quantization techniques to decrease the training model for GMM in order to reduce the processing time. VQ and GMM are widely applied to the speaker verification, but both have some disadvantages. To overcome those shortages, we introduce a new hybrid VQ/GMM model. Although in baseline form, the VQ-based solution is less accurate than the GMM, but it offers simplicity in computation. Therefore, we hope to make use of their merits via a hybrid VQ/GMM classifier.

5.3 Review of Previous VQ/ GMM Hybrid Methods

Previous studies have shown that the Vector Quantization techniques is insufficient to provide a high accuracy classification rate for speaker identification system if compare Gaussian mixture model. Nevertheless, VQ gain a good reputation of it simplicity and fast computation process. Consequently, most of the optimization speaker recognition systems use VQ technique to improve their baseline system.

There are many forms of GMM and other pattern classification techniques adaptation in the past. In hybrid VQ/GMM, most of them use VQ as optimization function to reduce the Expectation Maximization algorithm in order to improve the training speed [14]. Besides, some researchers use GMM as a post-processor after VQ cluster the speech signal into regions [15].

Gurmeet Singh et al. [16] introduced the use of two Vector Quantization algorithms, namely Linde, Buzo, Gray (LBG) and K-means algorithm for training Gaussian mixture speaker models as a replacement for Expectation Maximization algorithm to reduce computational complexity. However, if the speaker data become large, it still faces the time consuming problem.

Tomi Kinnunen et al. [17] presented another approach to optimizing vector quantization (VQ) based speaker identification. They do the pre-quantizing process to pruning out unlikely speakers; the best variants are then generalized to GMM based modeling. Based on their pruning idea, we propose using VQ techniques to make a decision rules before testing speaker data.

5.4Vector Quantization Decision Rules for Gaussian Mixture Modeling

Text independent speaker identification system requests a classifier that can classify a numbers of different data for input text by different speaker for testing phase. But usually, a big amount of training data and the difference of data, classified among varied classes take a long time processing. For GMM pattern classifier, it characteristic is to represent each speaker data into different GMM to generate a speaker model for training and testing. Even though it gain high accuracy rates, but it request a complex computation phase and long processing time.

For VQ, the primary factor is the codebook sizes [18], an experiment done by Kin Yu et al indicate that the optimum size is not dependent on the amount of training data. When a codebook is generated, its only remains the centroid which can represent the whole cluster. The amount of data is significantly less, since the number of centroids is at least ten times smaller than the number of vectors in the original sample. This will reduce the amount of computations needed when comparing in later stages. In fact, VQ based solution is less accurate than the GMM. Because VQ model and GMM model each one has advantages and disadvantages, in this paper, we make use of their merits, establishes VQ and the GMM mixture model for pattern classifier. In our proposed hybrid modeling, we take the superiority of VQ, which is simplicity computation to distinguish between male and female speaker. The purpose of this process is to divide the speaker into smaller subgroup.

The overall structure of our hybrid system is depicted in figure 11. After MFCC feature extraction process, the speech signal will transform to a feature vector form. For the phase 1 of the classification, VQ classifier clustering the speaker model into two sub-groups by decision tree structure. It is the male subgroup and the female subgroup. In phase 2 classification, we utilize dominance of GMM model to get the accuracy rates. GMM process will just applied in the particular subgroup to identify the speaker identity. GMM classification engine will calculate log likelihood score for subgroup 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. 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. Beside, it provides more simplicity in calculation[19,20].

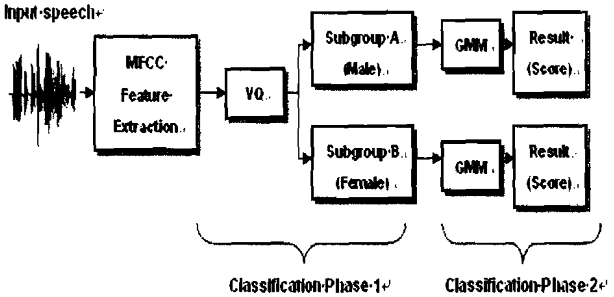

Figure. 11. Our propose method, speaker identification system based on vector quantization decision rules for Gaussian mixture modeling.

6.0 EXPERIMENTAL SETUP & RESULTS

In this section, we describe the experiments carried out in order to test the different recognizers as stated as above and make a comparison result with our hybrid technique. Experiments are conducted on a clean condition. In orders to get a fair comparison between 5 types of classifier, for each of then we have properly selected the same datasets and done some pro-processing for enhanced the feature data through a set of preliminary experiments.

6.1 Dataset Description

We performed our evaluation on the TIMIT speech database. The TIMIT corpus of read speech has been designed to provide speech data for development and evaluation of automatic speech recognition systems. However, the large number of distinct speakers present in the system also makes it suitable for evaluation speaker recognition system as well. TIMIT contains a total of 6300 sentences, 10 sentences spoken by 630 speakers from 8 major dialect regions of United Stated. Out of this large set, we chose 5 utterances of 10 distinct users to evaluate our system.

6.2 DTW System Evaluation

The first method evaluated uses DTW as pattern classification techniques. To evaluate the system, each sample utterance of the user was compared with the rest of the utterances in the database. For each comparison, the distance measure was calculated. A lower distance measure indicates a higher similarity. It is also of interest to see the effect increasing the number of speakers on the accuracy results besides comparison classifier performance. The first set of experiments; we use the TIMIT corpus only for these results increasing the number of speakers from 10 to 50.

Figure 12 shows the effect of increasing the speakers on performance of the DTW speaker identification system. Accuracy starts off highly 92% as would be expected, and slowly declines to approximately 80%. These results serve to show with increasing amounts of training data, the DTW distance measure become hard to calculated due to the progressively information of speaker.

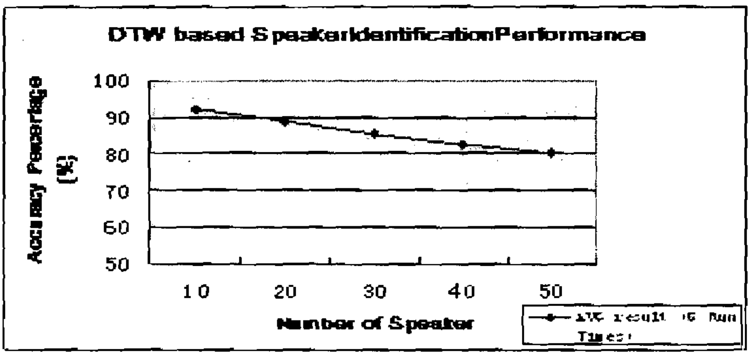

Figure 12. Performance of the DTW Speaker Identification System

6.3 GMM System Evaluation

The second method evaluated uses GMM as pattern classification techniques. Given training speech from a speaker's voice, the goal of this speaker model training is to estimate the parameter of GMM, which in some sense best matches the distribution of the training feature vector. We use maximum likelihood (ML) estimation in our experiment. The aim of ML is to find the model parameters, which maximize the likelihood of the GMM given the training data. Therefore, the testing data which gain a maximum score will recognize as speaker.

The second set of experiments; we use the TIMIT corpus only for these results increasing the number of speakers from 10 to 50. Figure 13 shows the effect of increasing the speakers on performance of the GMM speaker identification system. Accuracy starts off highly 98% as would be expected, and slowly declines to approximately 83%, which is congruent with accuracy results found by Reynolds [6]. As can be observed, even GMM speaker verification accuracy rate has decrease when the training data increase, but it still obtain the better result if compare with DTW.

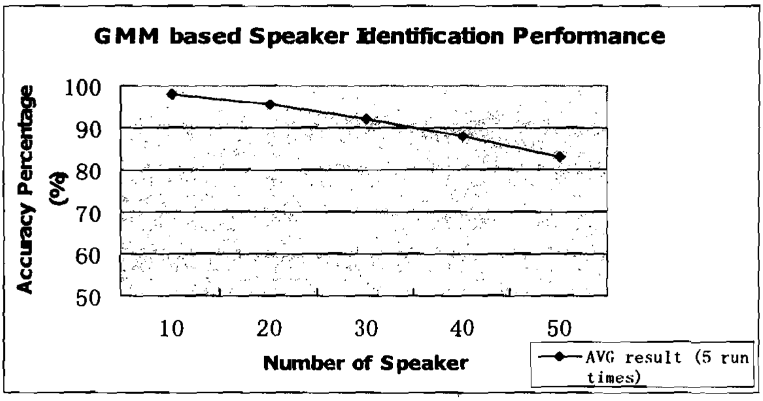

Figure 13. Performance of the GMM Speaker Identification System

6.4 SVM System Evaluation

The third method evaluated uses SVM as pattern classification techniques. An SVM is essentially a binary classifier trained to estimated whether an input vector x belongs to a class 1 (the desired output would be then y=+1) or to a class 2 (y=-1) where class 1 is verify as speaker and class 2 is verify as imposter.

The third set of experiments; we use the TIMIT corpus only for these results increasing the number of speakers from 10 to 50. Figure 14 shows the effect of increasing the speakers on performance of the SVM speaker identification system. Accuracy starts off highly 72% as would be expected, and slowly declines to approximately 60%. As can be observed, SVM gain the worse result if compare with DTW and GMM.

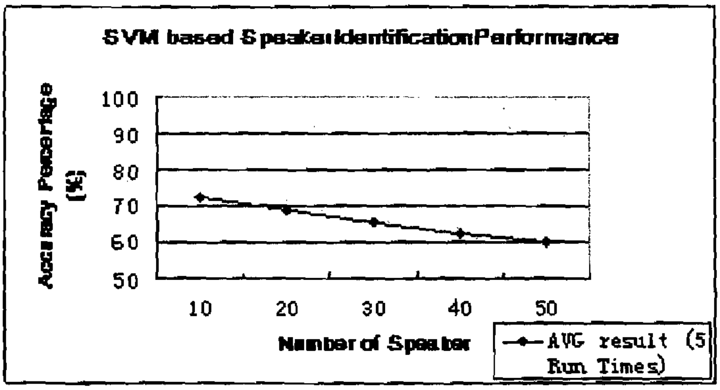

Figure 14. Performance of the SVM Speaker Identification System

6.5 VQ System Evaluation

VQ maps vectors to smaller regions called cluster. These cluster's center, centroid, are collected and will make up a codebook. The speaker identification engine are depends on the cookbook to identify a speaker. The fourth set of experiments; we use the TIMIT corpus only for these results increasing the number of speakers from 10 to 50. Figure 15 shows the effect of increasing the speakers on performance of the VQ speaker identification system. Accuracy starts off highly 96% as would be expected, and slowly declines to approximately 82%. As can be observed, VQ gain the worse result if compare GMM but still better then DTW and SVM.

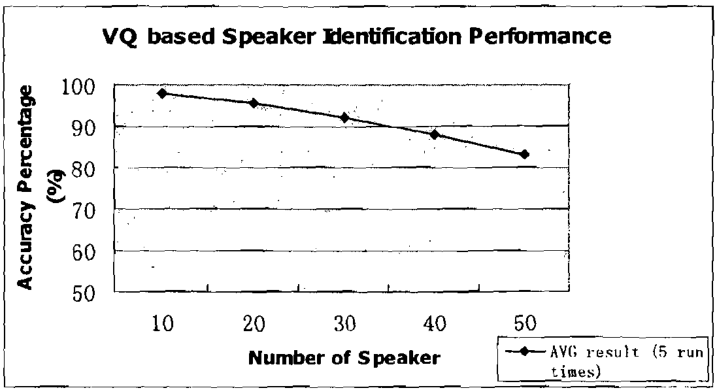

Figure 15. Performance of the VQ Speaker Identification System

6.6 Hybrid Vector Quantization/Gaussian Mixture Model System Evaluation

The last method evaluated uses hybrid VQ/GMM as pattern classification techniques. This is the new hybrid pattern classification as we proposed for speaker identification system. Here, we classified speaker by two phase of classification which the first phase we distinguish the male and female speakers using VQ decision approach and in the second phase of classification, GMM is applied into the subgroup of speaker. Figure 16 shows the effect of increasing the speakers on performance of the hybrid VQ/GMM speaker identification system for speakers from 10 to 50. Accuracy starts off highly 98.56%, and slowly declines to approximately 92.23%.

As can be observed, even hybrid VQ/GMM speaker identification accuracy rate has decrease when the training data increase, but it still obtain the better result if compare with other pattern classification method. Besides, it seems more stable to handle the large data set.

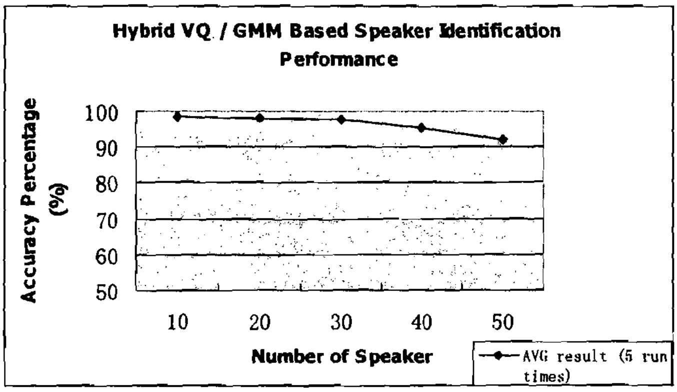

Figure 16. The performance of our propose method, hybrid vector quantization/Gaussian mixture model based speaker identification on increasing speaker data from 10-50.

6.7 Result for Processing Time

The result of time processing for 10 speakers by using GMM, VQ, DTW, SVM and hybrid VQ/GMM shows in table 1. We report that the GMM need 62.49 seconds for the whole training and testing process, VQ needs 38.71sec, DTW needs 54.12sec and SVM needs 96.13sec whereas our hybrid VQ/GMM just need 41.76 seconds. Thus, our implementation can categorized as more simplified version for classification techniques in speaker identification system. Obviously, a significant improvement compared to the baseline system is reported, a reduction in identification times up to 30% is reached.

Table 1. Comparison of time processing between 5 methods

Algorithm	GMM	VQ	DTW	SVM	VQ/GMM
Time	62.49sec	38.71sec	54.21sec	96.13sec	41.76sec
Number of speaker	10	10	10	10	10

6.8 Discussion

Results from experiments shows GMM likelihood function and VQ are well understood statistical model whereas DTW suitable due with small fixed vocabulary system. As can be observed, SVM gain the worse result among 4 types of classifier. This is due to the drawback of SVM when dealing with audio data is their restriction to work with fixed-length vectors. Obviously, the functions we choose for fixed-length vectors affect the performance of the SVM directly. However, among 5 types of pattern classification techniques, our hybrid techniques gain the best result in term of the accuracy rates and the processing time. Thus, our method provides an alternative way for real time identification system which time is the important issue.

7. CONCLUSION AND FUTURE WORK

In this paper, we have presented a model that is suitable for handle large datasets of speaker for speaker recognition system. Overall, the experiments have shown that the combination algorithm perform very efficiently and competitively on dataset under consideration. The algorithm compares well with published approaches and it is relatively easy to implement. Besides that, there is still a lot of enhancement can be make towards to this algorithm so that it can provide better outcomes. In conclusion, we have successfully improved the computation, approximation quality and accuracy of the speaker recognition system in this research.

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

[2] Sadaoki Furui., "Recent advances in speaker recognition", Pattern Recognition Letters. 1997,18 (9): 859-72.

[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] Lubkin, J. and Cauwenberghs, G., "VLSI Implementation of Fuzzy Adaptive Resonance and Learning Vector Quantization", Int. J. Analog Integrated Circuits and Signal Processing, vol. 30 (2), 2002,pp. 149-157.

[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] Temko, A.; Monte, E.; Nadeu, C., "Comparison of Sequence Discriminant Support Vector Machines for Acoustic Event Classification", ICASSP 2006 Proceedings, 2006 IEEE International Conference on Volume 5, Issue , 14-19 May 2006 Page 721-724.

[9] Shang, S.; Mirabbasi, S.; Saleh, R., "A technique for DC-offset removal and carrier phase error compensation in integrated wireless receivers Circuits and Systems", ISCAS apos;03. Proceedings of the 2003 International Symposium onVolume 1, Issue , 25-28 May 2003 Page I-173 - I-176 vol.1

[10] Vergin, R.; Oapos;Shaughnessy, D., "Pre-emphasis and speech recognition Electrical and Computer Engineering", Canadian Conference on Volume 2, Issue , 5-8 Sep 1995 Page 1062 - 1065 vol.2

[11] Davis, S. B. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustic, Speech and Signal Processing, ASSP-28, 1980, No. 4.

[12] Sadaoki Furui., "Cepstral analysis technique for automatic speaker verification", IEEE Trans. ASSP 29, 1981,pages 254-272.

[13] W.M. Campbell, J.P. Campbell, D.A. Reynolds, and E. Singer, "Support vector machines for speaker and language recognition," Computer Speech Lang., vol.20, no.2–3, April 2006, pp.210–229.

[14] 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.

[15] 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.

[16] 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.

[17] Kinnunen, Karpov and Franti, Real-Time Speaker Identification and Verification, IEEE Trans. On Audio, Speech and Language Processing, Vol. 14, No. 1, 2006.

[18] 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.

[19] Loh Mun Yee, Abdul Manan Ahmad and Fatimah Bte Mohamad, "Speaker Classification Optimization Design by Vector Quantization Decision Rule", ICTS 2008, 4th International Conference on Information & Communication Technology and Systems 2008 (Technical support by IEEE Indonesia Section), August 5th, 2008 in Surabaya, Indonesia.

[20] Loh Mun Yee and Abdul Manan Ahmad, "Text-Independent Speaker Identification Using Hybrid Vector Quantization / Gaussian Mixture Models Pattern Classifier ",ICCAS 2008, IEEE Computer Society Press in the Proceedings of the International Conference on Control, Automation and Systems 2008, Oct. 14-17,2008 in Seoul, Korea.

[21] Loh Mun Yee, Abdul Manan Ahmad and Cheang Soo Yee, "Selecting Appropriate Feature Extraction Techniques For Hybrid Pattern Classification Speaker Identification Modeling",ICPE-3 2008, 3rd International Conference On Postgraduate Education, December 16-17, 2008, Gurney Hotel, Penang, Malaysia , in press.