The 5th Student Conference on Research and Development –SCOReD 2007
11-12 December 2007, Malaysia

# Speaker Verification using Vector Quantization and Hidden Markov Model

Mohd Zaizu Ilyas, *Member, IEEE*, Salina Abdul Samad, *Senior Member, IEEE,* Aini Hussain, *Member , IEEE* and Khairul Anuar Ishak, *Member, IEEE*

*Abstract*-- **This paper presents a speaker verification system using a combination of Vector Quantization (VQ) and Hidden Markov Model (HMM) to improve the HMM performance. A Malay spoken digit database which contains 100 speakers is used for the testing and validation modules. It is shown that, by using the proposed combination technique, a total success rate (TSR) of 99.97% is achieved and it is an improvement of 11.24% in performance compared to HMM. For speaker verification, true speaker rejection rate, impostor acceptance rate and equal error rate (EER) are also improved significantly compared to HMM.**

***Index Terms*-- Speaker recognition, speaker verification, hidden Markov model, vector quantization**

## I. Introduction

SPEAKER or voice recognition or verification is a biometric modality that uses an individual's voice for recognition or verification purpose. It is a different technology from speech recognition, which recognizes words as they are articulated, which is not biometrics[10]. Speech contains many characteristics that are specific to each individual. For this reason, listeners are often able to recognize the speaker's identity fairly quickly even without looking at the speaker. Speaker verification is a process of determining whether a person is who he or she claims to be by using his or her voice [5,6,8,9,10].

There are two types of speaker verification, text dependent and text independent. In text dependent speaker verification, a speaker presents fixed or prompted phrase that is programmed into the system and can improve system performance. In a text independent speaker verification system, the system has no advance knowledge of the speaker's phrasing, and is much more difficult and less robust [5,6,10]. Text dependent speaker verification is currently the most commercially viable and useful technology although both systems have been rapidly researched. Speaker verification has many potential applications, including access control to computers, databases and facilities, electronic commerce, forensic and telephone banking.

For many years research on speaker verification has been done and some of them have reached high performance level. Many techniques have been proposed for speaker verification systems including dynamic time wrapping (DTW), Hidden Markov models (HMM), artificial neural networks (ANN) and vector quantization (VQ) [6]. Recent studies show that high performance for text dependent speaker verification can be achieved using HMM approach [5, 10]. This paper presents a text dependent speaker verification using a combination approach of VQ and HMM. The objective is to improve the performance of HMM in a speaker verification system. The proposed technique is evaluated using Malay spoken digit database which contains 100 speakers obtained in noise-free environment. The results are compared with stand alone HMM. The remaining part of this paper is organized as follows. Section II and section III present the details of VQ and HMM technique. Section IV describes the Malay spoken digit database. Experiments and results are discussed in sections V and VI. Finally, concluding remarks are presented in section VII.

## II. Vector Quantization

Vector Quantization (VQ) is a process of mapping vectors of a large vector space to a finite number of regions in that space. Each region is called a *cluster* and is represented by its centre (called a *centroid*) [3]. A collection of all the centroids makes up a codebook. 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 for comparison in later stages. Even though the codebook is smaller than the original sample, it still accurately represents a person's voice characteristics. The only difference is that there will be some spectral distortion.

In an earlier feature extraction stage, we calculate the LPC cepstrum, and the entire speech signal are represented as the LPC to cepstrum parameters and a large sample of these parameters are generated as the training vectors. During the training process of VQ, a codebook is obtained from these sets of training vectors. These training vectors are actually compressed to reduce the storage requirement. An element in a finite set of spectra in a codebook is called a codevector. The codebooks are used to generate indices or discrete symbols

This research is supported by the following research grant: Fundamental Research Grant Scheme, Malaysian Ministry of Higher Education, FRGS UKM-KK-02-FRGS-0036-2006

M.Z. Ilyas is a M.Sc. student at Department of Electrical, Electronic & System Engineering, Universiti Kebangsaan Malaysia, 43600, UKM, Bangi, Malaysia (e-mail: mozazy@vlsi.eng.ukm.my ).

S.A. Samad and A. Hussain are professors and K.A. Ishak is a lecturer at Department of Electrical, Electronic & System Engineering, Universiti Kebangsaan Malaysia, 43600, UKM, Bangi, Malaysia (e-mail: salina@vlsi.eng.ukm.my, aini@vlsi.eng.ukm.my & nuarscc@vlsi.eng.ukm.my ).

that will be used by the discrete HMM. Hence, data compression of speech is accomplished by VQ in the training phase and the encoding phase that finds the input vectors the best codevectors.

To implement VQ, first, we must get the codebook. A large set of spectral analysis vectors (or speech feature vectors) is required to form the training step. If we denote the size of the VQ codebook as $M = 2^N$ codewords, then we require an L (with L >> M) number of training vectors [2]. It has been found that L should at least be 10M in order to train a VQ codebook that works well. For this project, we use the LBG algorithm, also known as the binary split algorithm.

## III. HIDDEN MARKOV MODEL

A speaker verification system consists of two phases which is the training phase and the verification phase. In the training phase, the speaker voices are recorded and processed in order to generate the model to store in the database. While, in the verification phase, the existing reference templates are compared with the unknown voice input. In this project, we use the Hidden Markov Model (HMM) method as the training/recognition algorithm.

The most flexible and successful approach to speech recognition so far has been HMM. The goal of HMM parameter estimation is to maximize the likelihood of the data under the given parameter setting. General theory of HMM has been given in [2,7,8]. There are 3 basic parameters in HMM which is:

- $\boldsymbol{\pi}$ - The initial state distribution.
- **a** – The state-transition probability matrix.
- **b** – Observation probability distribution.

In the training phase, a HMM model for each speaker is generated. Each model is an optimized model for the word it represents. For example, a model for the word 'Satu' (number one), has its **a**, **b**, and $\boldsymbol{\pi}$ parameters adjusted so as to give the highest probability score whenever the word 'Satu' is uttered, and lower scores for other words. Thus, to build a model for each speaker, a training set is needed. This training set consists of sequences of discrete symbols, such as the codebook indices obtained from the Vector Quantization stage.

Here, an example is given of how HMM is used to build models for a given training set. Assuming that N speakers are to be verified, first we must have a training set of L token words, and an independent testing set. To do speaker verification, the following steps are needed:

1. First we build an HMM for each speaker. The L training set of tokens for each speaker will be used to find the optimum parameters for each word model. This is done using the re-estimation formula.
2. Then, for each unknown speaker in the testing set, first characterize the speech utterance into an observation sequence. This means using an analysis method for the speech utterance so that we get the feature vector, and then the vector is quantized using Vector Quantization. Thus, we will get a sequence of symbols, with each symbol representing the speech feature for every discrete time step.
3. We calculate **a**, **b** and $\boldsymbol{\pi}$ parameters for the observation sequence using one of the speaker models in the vocabulary. Then repeat for every speaker model in the database.

After N models have been created, the HMM engine is then ready for speaker verification. A test observation sequences from an unknown speech utterance (produced after vector quantization of cepstral coefficient vectors), will be evaluated using the Viterbi algorithm. The log-Viterbi algorithm is used to avoid precision underflow. For each speaker model, probability score for the unknown observation sequence is computed. The speaker whose model produces the highest probability score and matches the ID claimed is then selected as the client speaker.

Speaker verification means making a decision on whether to accept or reject a speaker. To decide, a threshold is used with each client speaker. If the unknown speaker's maximum probability score exceeds this threshold, then the unknown speaker is verified to be the client speaker (i.e., speaker accepted). However, if the unknown speaker's maximum probability score is lower than this threshold, then the unknown speaker is rejected. The relationship is shown in Figure 1.

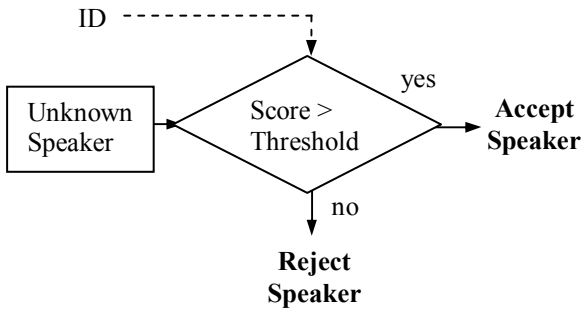


Fig. 1. Speaker verification decision.

The threshold is determined as follows:

1. For each speaker, evaluate all samples spoken by him using his own HMM models and find the probability scores. From the scores, find the mean, $\mu_1$, and standard deviation, $\sigma_1$, of the distribution.
2. For each speaker, evaluate all samples spoken by a large number of impostors (typically over 20) using the speaker's HMM models and find the probability scores. From the scores, find the mean $\mu_2$ and standard deviation $\sigma_2$ of the distribution.
3. For each speaker, calculate the threshold as:

$$T = \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1 + \sigma_2} \quad (1)$$

## IV. MALAY SPOKEN DIGIT DATABASE

The raw Malay Spoken digit database was collected at Faculty of Language and Linguistic, University Malaya as part of a Malay corpus database. The database was analyzed, processed at the Signal Processing Laboratory, Faculty of Engineering, Universiti Kebangsaan Malaysia. The Malay spoken digit database contains continuous spoken digit from 0 to 9 in slow (with silent gaps or stop) and fast (without silent gaps or stops) speech and obtained in recording room environment. The database comprises of 212 Mb of spoken digit speech spoken by 100 speakers of different races, ages and background. The speech material is stored in wav format with 16-bit of radio sample size and at 16 KHz audio sample rate. Table 1 summarizes the database. Out of 100 speakers, 16 of them are males and 84 of them are females. The highest population of the speakers was Malay which represents about 49% of the population. Chinese represents 35% of the population followed by Indian of about 13%. The average speaker age is about 26 years old and represents more than a half of the total populations.

TABLE 1
MALAY SPOKEN DIGIT DATABASE

| | |
|---|---|
| Speakers | 100 ( 16 Male / 84 Female) |
| Session/Speakers | 1 |
| Type of speech | Prompted Malay digit (0 to 9) |
| Microphone | Standard microphone |
| Acoustic environment | Recording room (±55dB) |
| Audio sample size | 16 bit |
| Audio sample rate | 16 KHz |
| File format | Wave |

## V. EXPERIMENTS

### A. Experimental Setup

Speaker verification experiments were carried out using the database described in [IV]. 100 speakers were selected where each speaker has 10 repetitions of Malay digits. All Malay digits, from 0 until 9, were selected to build the speaker model. The samples were divided into 2 sets, one for training session and the other for the testing session. Feature vectors comprising of 14 cepstrum coefficients computed using Linear Prediction Cepstral Coefficient (LPCC) [3] were used. The analyzed frame was windowed by a 15 milliseconds Hamming window with 5 milliseconds overlapping. All samples were down-sampled to 16 kHz prior to feature extraction. Speaker verification using HMM approach and combination of VQ and HMM approach was tested. All experiments were conducted using a desktop computer, equipped with a Pentium D processor, with 1 GBytes of memory and running on the Windows XP operating system.

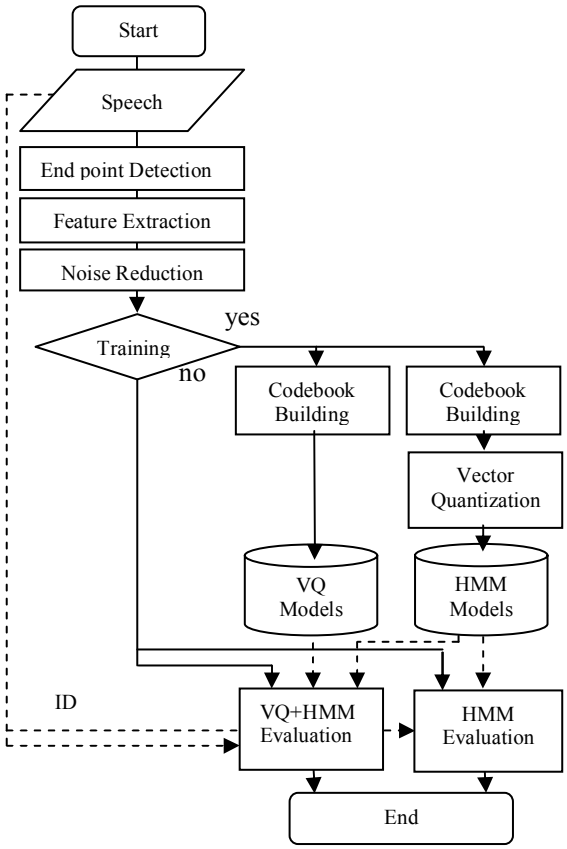


Fig. 2. Speaker verification experiment flow chart

Fig. 2. shows the flow chart of the speaker verification experiment. First, speech signal will go through end point detection, feature extraction and noise reduction. If it is a training process it will generate individual codebook for VQ models and global codebook for HMM models. Otherwise, evaluation of combination VQ and HMM and standalone HMM will be conducted based on individuals models and ID.

### B. Performance Criteria

The basic error measures of a verification system are false acceptance rate (FAR) and false rejection rate (FRR), as defined in (2) and (3).

$$\text{FAR} = \frac{\text{Number of accepted imposter claims}}{\text{total number of imposter accesses}} \text{ x } 100 \quad (2)$$

$$\text{FRR} = \frac{\text{Number of rejected genuine claims}}{\text{total number of genuine accesses}} \text{ x } 100 \quad (3)$$

Overall performance can be obtained by combining these two errors into total success rate (TSR) where:

$$\text{TSR} = 100\% - \left( \frac{\text{FAR} + \text{FRR}}{\text{Total number of accesses}} \right) \times 100 \quad (4)$$

Speaker verification threshold or equal error rate (EER) is calculated as (1).

## VI. RESULTS AND DISCUSSION

Table 2 shows a summary of the verification results for the experiments performed. A total success rate (TSR) of 99.97% was achieved using this combination technique compared to stand alone HMM which is 89.87%. The TSR performance improve significantly by 11.24%. Using the combination technique, true speaker rejection rate achieved was 0.06% while impostor acceptance rate was 0.03% and equal error rate (EER) of 11.72% was achieved. Figure 3 shows a ROC plot of False Rejection Rate(FRR) vs False Acceptance Rate (FAR). It shows that a combination technique between VQ and HMM outperformed the HMM based technique.

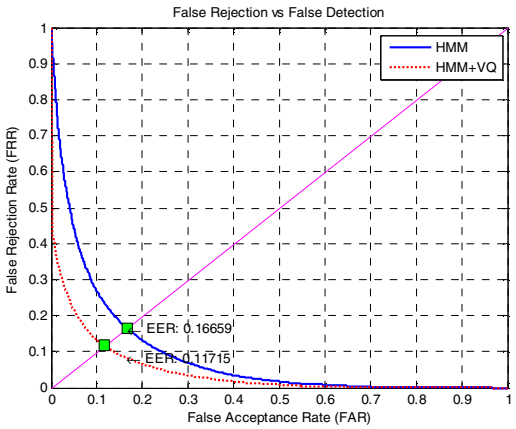


Fig. 3. ROC plot of False Rejection Rate(FRR) vs False Acceptance Rate (FAR)

TABLE 2
MALAY SPOKEN DIGIT DATABASE

| Method | FRR | FAR | TSR | EER |
|---|---|---|---|---|
| HMM | 25.30% | 9.99% | 89.87% | 16.66% |
| VQ+HMM | 0.06% | 0.03% | 99.97% | 11.72% |

## VII. CONCLUSION

This paper has shown that the combination approach of VQ and HMM can improve the HMM performance in a noise-free environment. The Malay spoken digit database which contains 100 speakers has been used to test and validate the system. It is shown that, a total success rate (TSR) of 99.97% was achieved using this combination technique compared to HMM which was 89.87%. The TSRs performance improves significantly by 11.24%. For FRR, FAR, and EER, the combination technique also shows improvement. Further work will concentrate on noisy environment to evaluate the robustness of the system.

## VIII. REFERENCES


[1] A. Peinado and J. C. Segura, Speech Recognition over digital channel:robustness and standards, John Wiley and Sons, England, 2006.

[2] C. Wheddon and R. Linggard, *Speech and Language Processing*, Chapman and Hall, UK, 1990, pp. 209-230.

[3] F.K. Soong, A.E. Rosenberg, L.R. Rabiner and B.H. Juang, "A Vector Quantization approach to Speaker Recognition", Florida: ICASSP Vol.1, 1985, pp. 387-390.

[4] J.L. Wayman, "Error Rate Equations for the General Biometric System". *IEEE Robotic & Automation.*, 6(9), March 1999, pp.35-48.

[5] J.M. Naik, "Speaker Verification: A Tutorial", *IEEE Communication Magazine*, January 1990, pp. 42-48.

[6] J.P. Campbell, "Speaker Recognition: A tutorial ", *Proc. of the IEEE,* Vol. 85, No. 9, September 1997, pp. 1437 – 1462.

[7] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", *Proceeding of The IEEE*, Vol.77, No.2, February 1989.

[8] L.R. Rabiner and B.H. Juang, *Fundamental of Speech Recognition*, Prentice Hall, New Jersey, 1993.

[9] National Science and Technology Council (NTSC), "Speaker Recognition", 2006. [Online]. Available: http://www.biometricscatalog.org/NSTCSubcommittee/Documents/Speaker%20Recognition.pdf [7 August 2006].

[10] T. Matsui and S. Furui, "Comparison of Text Independent Speaker Recognition Methods using VQ-Distortion and Discrete/Continuous HMMs", *Proceedings of ICASSP-92*, Vol. 2, 1992, pp. 157-160.


## IX. BIOGRAPHIES

**Mohd Zaizu Ilyas** (M'2006) received an associate degree in electrical engineering from Tsuyama National College of Technology Japan in 1998 and B.Sc degree in electronics and information engineering from Tokyo University of Agriculture and Technology Japan in 2000.

From 2000 to 2006, he worked at ROHM Wako electronics (M) Co., as a senior engineer and was appointed as head of production engineering department in 2001 and head of instrumentation department in 2003. In July 2006, he joined Digital Signal Processing Research Group, at the Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering, Universiti Kebangsaan Malaysia (UKM), Malaysia. He currently is pursuing a M.Sc. degree in UKM. His research interests include speech recognition, speaker verification and digital filter design.

**Salina Abdul Samad** (M'1995-SM'2007) received the BSEE and Ph.D. in electrical engineering from University of Tennessee and University of Nottingham, respectively. She is a professor at the Department of Electrical, Electronic & Systems Engineering, Universiti Kebangsaan Malaysia. Her research interests include digital signal processing, filter design and multimodal biometrics.

**Aini Hussain** (M'1997) received the BSEE, M.Sc. and Ph.D. in electrical engineering from Louisiana State University, UMIST and Universiti Kebangsaan Malaysia, respectively. She is a professor at the Department of Electrical, Electronic & Systems Engineering, Universiti Kebangsaan Malaysia. Her research interests include intelligent signal processing, pattern recognition and system modeling.

**Khairul Anuar Ishak** (M'2007) received a B.Sc. in computer engineering from Universiti Teknologi Malaysia in 2001 and M.Sc. in electrical engineering from Universiti Kebangsaan Malaysia in 2006. He is currently a lecturer at the Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering, Universiti Kebangsaan Malaysia. His research interests include artificial intelligence, image processing, speech processing and pattern recognition.