Malay Speaker Recognition System Based On Discrete HMM

A. K. Ariff, M. Alwi, Sh-Hussain, Salleh

Centre for Biomedical Engineering
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 Skudai, Johor

E-mail: amarism@yahoo.com, ikesela@yahoo.com , hussain@fke.utm.my

Abstract

This paper presents the design and implementation of Malay speaker recognition system using discrete hidden Markov model (HMM) as the classifier. A series of speaker recognition experiments was performed using 99 speakers (13 clients and 86 imposters) recording database consisting of isolated digit utterances. For a seven digit long sequence, 0.96% EER was achieved.

1. Introduction

Speaker recognition is the process of automatically recognizing the person speaking on the basis of the information obtained from the speech features. It has been an important subject for research, and its development has come to a stage where it has been actively and successfully applied, especially in biometric applications.

Speaker recognition can be classified into two different categories : speaker verification and speaker identification. In speaker identification, there is no a priori identity claim, and the system decides who the person is, what group the person is a member of, or that the person is unknown. In a speaker verification task, the recognizer is asked to verify an identity claim made by an unknown speaker and a decision to reject or accept the identity claim is made [1].

In this paper, we described research on the speaker recognition using Malay digits. We present the results of applying Hidden Markov Modeling (HMM) as the recognition engine of the system. It is the state-of-the-art of various speaker recognition systems available today.

HMM have a number of very powerful properties [2]. The ability of HMM to automatically optimize parameters from data is extremely powerful, the HMM integrated search that considers all of the knowledge sources at every step is very effective, and the absorption of faulty structural assumptions is most forgiving. By turning an unknown structure problem into an unknown parameter problem, and by automatically optimizing these parameters, HMM and maximum likelihood estimation are one of the most powerful learning paradigms.

2. Speech Database

In this work, the experiment has been done on the limited training data, which only five samples of speech for each Malay digit (zero to nine) for each session was used. There is a total of five sessions of speech samples collected at different time in office environment. The distribution in time of recording session can be seen in Figure 1.

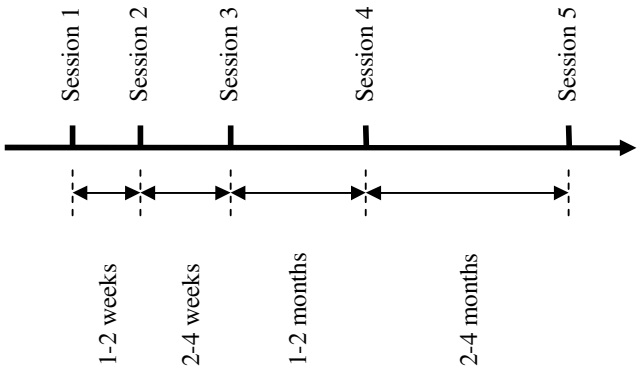

Figure 1 – Distribution In Time Of Recording Sessions

The recording session is taken in range of 1 to 2 weeks after the first session. In the third recording session the range is increased from 2 to 4 weeks; 1 to 2 months for fourth recording session and the fifth recording session is from 2 – 4 months range. Approximately 6 months time has been spend for speech recording session taken from 13 speakers. There was a total of 250 utterances for each of the speakers.

1-4244-0011-2/05/$20.00 ©2005 IEEE.

The database also consists of utterances of 86 speakers that were used for testing. Each speaker spoke 50 isolated digits (5 utterances per digit) in one session.

Speech recordings were recorded in 16 bit, 16kHz, raw format using high quality microphone through personal computer.

3. Speech Signal Processing

Most of today's automatic speech recognition systems are based on some type of Mel Frequency Cepstrum Coefficients (MFCCs), which have been proven to be effective and robust under various conditions. It provides good performance, better accuracy and minor computational complexity with respect to alternative features [3].

Therefore, MFCC analysis was performed on the sampled speech input. The frame blocking size was 240 points with 80 points overlap. After pre-emphasis (factor 0.95) and application of a Hamming window, 24 MFCC coefficients were computed.

4. Hidden Markov Model

Discrete density HMMs can be defined by :

- $\{s\}$ – a set of states sequence including an initial state S_i and a final state S_f
- $\{\pi\}$ – a set of probability of the first state
- $\{a_{ij}\}$ – a set of transitions where a_{ij} is the probability of taking a transition from state i to state j
- $\{b_{ij}(k)\}$ – the output probability matrix; the probability of emitting symbol k when taking a transition from state i to state j

The compact notation for hidden Markov model is $\lambda=(a,b,\pi)$.

For every digit, 5-state HMM, non-ergodic model were built, as shown in Figure 2.

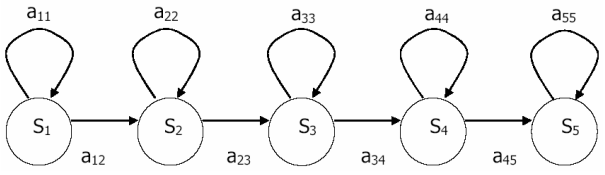

Figure 2 – The Hidden Markov Model

5. Experimental Setup and Results

Several experiments were carried out to evaluate the performance of the system. The effects, on performance, of different system parameters such as codebook sizes, digit sequence length and the number of recording sessions were also studied.

Effects of Vector Quantization Codebook Size

The speaker identification equal error rate (EER) is plotted as a function of codebook size in Figure 3. The codebook size is represented by the number of vector entries, M or by the corresponding rate, $M = 2^R$.

The EER decreases when the codebook size increases from 12.4% to 3.21%. This can be explained as the codebook size increases, the distortion (quantization) error decreases. The codebook with size of 256 can perform better than codebook of size 128 (3.74%), but the average time to recognize one speaker is higher with codebook size of 256, compared to codebook size of 128. This can affect the time response of identification system. Therefore, codebook size of 128 was used throughout the experiment.

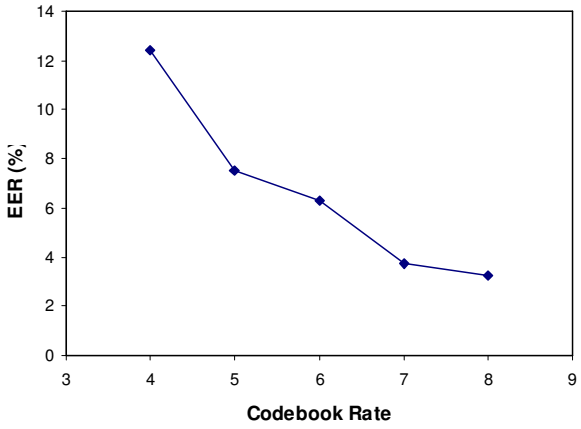

Figure 3 – EER Versus Codebook Rate

Single Digit Performance

The speaker identification results obtained using 10 single digits are shown in Figure 4. The digit 3 achieved the best results while the digits 4, 5 and 7 achieved the worst scores. While digit 1, 2 and 6 achieved a good results.

The average EER of a single digit will decrease as more utterances were recorded in many sessions. The EER plotted as a function of the recording session number

is shown in Figure 5.

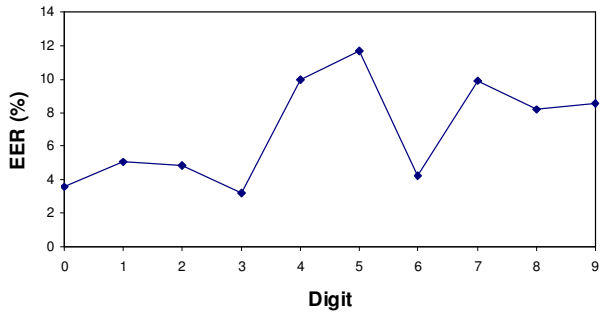

Figure 4 – Single Digit Performance

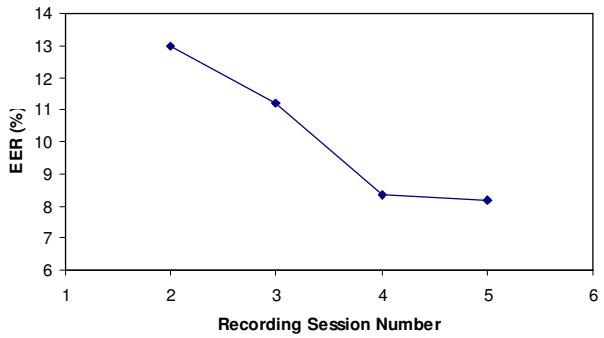

Figure 5 – Single Digit Performance

More training data will improve the performance of speaker recognition system as it has enough parameters to model variability in the data. By recording the utterances in more sessions, the codebook will be updated with new data. This step is very crucial as the intra-speaker variations will degrade the performance. The longer the separation between the training (recording session) and the test recording, the worse the performance.

Digit Sequence Performance

By increasing the length of digit in sequence, more discriminating information is given. The longer the utterance will contains more information for speaker discrimination. In [4] also shows that the performance is improved with the length of text to be spoken. Table 1 shows the performance over digit sequence speaker verification.

Figure 6 shows that as more and more digit being added, the performance is better than using single digit. The overall performance had an average EER of 8.18% with single digit sequence, while best performance with digit length of 8 giving an average EER of 1.67%.

Table 1 : Digit Sequence Performance

Digit Length	EER (%)
1	8.18
2	4.92
3	3.86
4	3.34
5	2.73
6	2.36
7	1.93
8	1.67
9	1.95
10	1.74

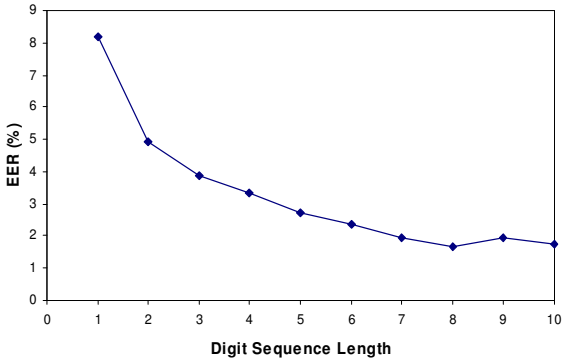

Figure 6 – Digit Sequence Performance

Selected Digit Sequence Performance

As described in single digit experiment, the performance of each digit varies from each other. In order to achieve better performance, the digit with poor performance should be removed. In this experiment only 7 digits with good performances have been selected after removing digit 4, digit 5 and digit 7. The overall performance is listed in Table 2, while graphical representation is shown in Figure 7.

Table 2 : Selected Digit Sequence Performance

Digit Length	EER (%)
1	8.18
2	4.92
3	3.86
4	3.34
5	2.23
6	1.54
7	0.96

The result in Figure 6 shows that using selected digit, provide better performance with digit length of 7, the average EER of 0.96% was achieved. There is an

improvement of 43% using the best digit selected when compared to the digit sequence from Table 1.

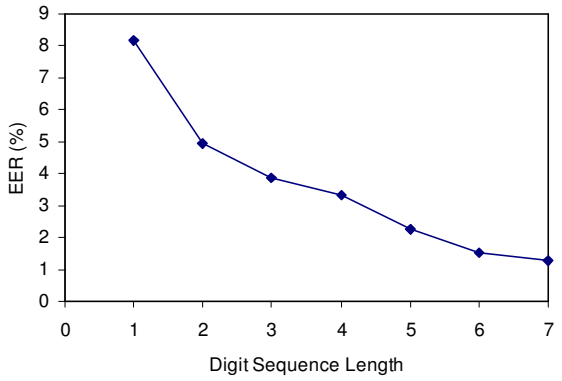

Figure 6 – Selected Digit Sequence Performance

6. Discussion

Several experiments were carried out to evaluate the performance of the speaker recognition system. The overall best performance is based on a 7 digit long sequence and a codebook of 128 vectors, where EER of 0.96% was achieved.

It was also shown that both larger codebook size and longer digit sequence (more digits in the test utterance) can be used to improve recognition performance. The codebook should always be updated from time to time to alleviate the performance degradation due to different recording conditions and intra-speaker variations.

An improvement in front-end analysis will also increase the performance of the system. It was suggested that differential information (delta- and delta-delta) and power information should be use in front-end process. It has been shown that the use of these information is extremely important [5].

7. Conclusion

We have designed a HMM-based speaker recognition and evaluated its performance on Malay digits. The system gave good results, even though there are much more room for improvement.

References

[1] Soong F.; Rosenberg A.; Rabiner L.; Juang B. 1985. A Vector Quantization Approach To Speaker Recognition. International Conference On Acoustics, Speech and Signal Processing, 387-390.

[2] Lee K. F. 1989. Hidden Markov Models : Past, Present and Future. In Proceedings of the European Conference on Speech Communication and Technology, 148-155. Paris.

[3] Davis S.B.: Mermelstein P. 1980. Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans. Acoustic, Speech and Signal Processing, 357-366. .

[4] Ariyaeeinia A. M.; Sivakumaran P., 1997. Comparison of VQ and DTW Classifiers for Speaker Verification. European Conference on Security and Detection, No. 437, 142-145.

[5] Furui S.. 1986. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, IEEE Transaction on Acoustic, Speech and Signal Processing, 52-59.