American Journal of Applied Sciences 5 (6): 714-720, 2008
ISSN 1546-9239

Isolated Malay Digit Recognition Using Pattern Recognition Fusion of Dynamic Time Warping and Hidden Markov Models

S.A.R. Al-Haddad, S.A. Samad, A. Hussain, K.A. Ishak
Department of Electrical, Electronic and System Engineering, Faculty of Engineering,
University Kebangsaan Malaysia, 43600, UKM, Bangi, Malaysia

Abstract: This paper is presents a pattern recognition fusion method for isolated Malay digit recognition using Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). The aim of the project is to increase the accuracy percentage of Malay speech recognition. This study proposes an algorithm for pattern recognition fusion of the recognition models. The endpoint detection, framing, normalization, Mel Frequency Cepstral Coefficient (MFCC) and vector quantization techniques are used to process speech samples to accomplish the recognition. Pattern recognition fusion method is then used to combine the results of DTW and HMM which uses weight mean vectors. The algorithm is tested on speech samples that are a part of a Malay corpus. This paper has shown that the fusion technique can be used to fuse the pattern recognition outputs of DTW and HMM. Furthermore it also introduced refinement normalization by using weight mean vector to get better performance with accuracy of 94% on pattern recognition fusion HMM and DTW. Unlikely accuracy for DTW and HMM, which is 80.5% and 90.7% respectively.

Key words: Endpoint detection, Mel frequency cepstral coefficient, vector quantization, distance measurements and weight mean

INTRODUCTION

In many speech recognition systems, endpoint detection and pattern recognition are used to detect the presence of speech in a background of noise. The beginning and end of a word should be detected by the system that processes the word. The problem of detecting the endpoints would seem to be easily distinguished by human, but it has been found complicated for machine to recognize. Instead in the last three decades, a number of endpoint detection methods have been developed to improve the speed and accuracy of a speech recognition system. This study uses the Malay language, which is a branch of the Austronesian (Malayo-Polynesian) language family, spoken as a native language by more than 33,000,000 persons distributed over the Malay Peninsula, Sumatra, Borneo, and the numerous smaller islands of the area, and widely used in Malaysia and Indonesia as a second language [1].

Speech Recognition (SR) is a technique aimed at converting a speaker's spoken utterance into a text string or other applications. SR is still far from a solved problem. It is quoted that the best reported word-error rates on English broadcast news and conversational telephone speech were 10% and 20%, respectively [2]. Meanwhile, error rates on conversational meeting speech are about 50% higher, and much more under noisy conditions [3].

This paper proposes a fusion pattern recognition method for isolated Malay digit recognition using Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). DTW was used in speech recognition in 70's and 80's[4, 5] and HMM was popular after 90's and still continues now [6]. Meanwhile fusion techniques are used to solve biometric problems especially for sensors, extractors, classifiers and supervisors as shown in Fig. 1[7]. For further improvement in speech recognition [8] used a technique, which used decision fusion for making decisions as shown in Fig. 2.

In Figure 2, it shows the Confidence Measure (CM) as assessing the reliability of recognition results in two ways: Fig. 2(a) Feature-level fusion and Fig. 2(b) Decision-level fusion. Furthermore CM gives a confident estimation and is followed by a decision whether to reject or accept the hypothesized isolated

Corresponding Author: Syed Abdul Rahman Al-Haddad, Signal Processing Research Group, Department of Electrical, Electronic and System Engineering, Universiti Kebangsaan Malaysia, 43600, UKM, Bangi, Malaysia
Tel: +603-89216317 Fax: + 603-89216146

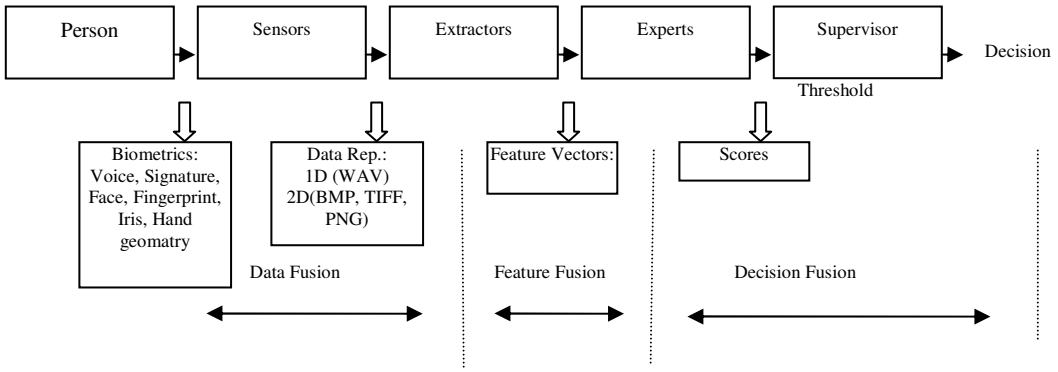

Fig. 1: A generic biometric taxonomy and fusion scheme

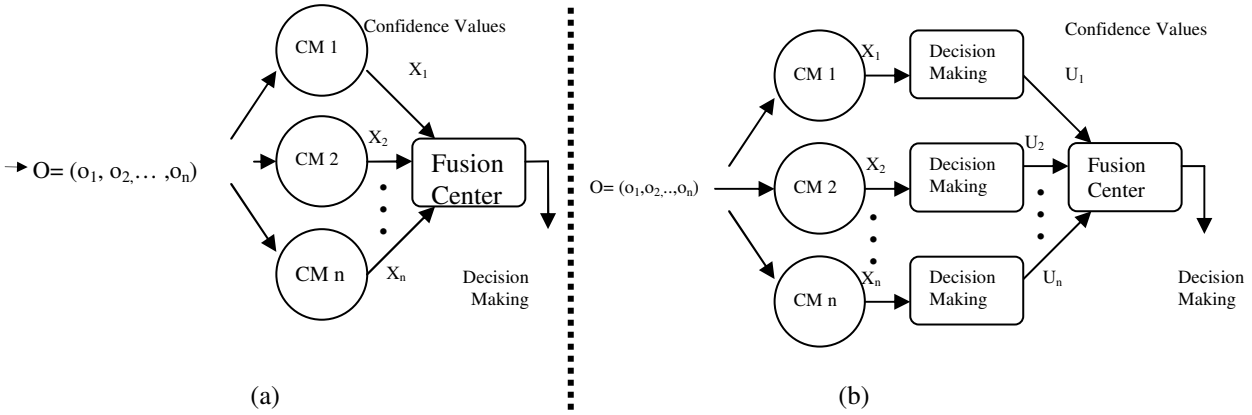

Fig. 2: Two types of confidence measure fusion (a) Feature-level fusion (b) Decision-level fusion

digit. In this paper the decision-level fusion is used where decision making X1 is assumed as DTW, decision making X2 as HMM, and the fusion centre or pattern recognition fusion method as weight mean vector is as shown in Fig. 3.

The aims of this project are: (i) to increase the accuracy percentage of Malay speech recognition; (ii) to develop patterns of reference for the Malay digit in the recognition database by using HMM and DTW; and (iii) fuse them using weight mean vector for improving the recognition.

This paper is segmented in 4 sections: Introductions, materials and Methods, Results and Discussion and Conclusions.

MATERIALS AND METHODS

The algorithm is tested on Malay digit speech corpus. The Malay isolated digits are from 0 to 9 spoken as KOSONG, SATU, DUA, TIGA, EMPAT, LIMA, ENAM, TUJUH, LAPAN and SEMBILAN with 10 repetitions for each digit. The system begins with the input speech, end point detection, framing, normalization, filtering, MFCC, time normalization, and using HMM and DTW to calculate the reference patterns. The DTW is used to normalize the training data with the reference patterns respectively as shown in Fig. 3. Finally weighted mean vector is used with the results from DTW and HMM to get the final decision output.

End Point Detection: After getting the speech sample, the first process is endpoint detection. For detection, two basic parameters are used: Zero Crossing Rate (ZCR) and short time energy. The energy parameter has been used in endpoint detection since the[9]. By combining with the ZCR, speech detection process can be made very accurate [10]. The begin and end for each utterance can be detected.

The measurements of the short time energy can be defined as follows [11]:

a. logarithm energy:

$$E = \sum_{i=1}^{N} \log x^2(i) \quad (1)$$

b. sum of square energy:

$$E = \sum_{i=1}^{N} x^2(i) \quad (2)$$

c. sum of absolute energy:

$$E = \sum_{i=1}^{N} | x^2(i) | \quad (3)$$

As mentioned in the definition above, we write the algorithm E as energy, N as samples in a frame. The frame size is 256, sample rate is 8 KHZ, the upper level energy is -10dB and the lower level energy is -20dB.

The flowchart of the end point detection is shown on Fig. 4. The system begins with reading a WAV file, which is recorded, from 15 male and 15 female speakers. Each speaker says KOSONG, SATU, DUA, TIGA, EMPAT, LIMA, ENAM, TUJUH, LAPAN and SEMBILAN with a 1 second of pause between each digit.

Then the ZCR is adjusted to the number of times in a sound sample the amplitude of the sound wave changes the sign by getting their mean. A tolerance of threshold is included in the function that calculates zero crossing which is 10% of maximum ZCR. Next logarithm (log) energy allows us to calculate the amount of energy in a sound at specific instances. For specific window size there are no standard values of energy. Log energy depends on the energy in the signal, which changes depending on how the sound was recorded. In a clean recording of speech the log energy is higher for voiced speech and zero or close to zero for silence.

It expands the endpoint lower level by reversing the sound index until it reaches the first point's energy which falls below a low level energy threshold. Next it expands the end point for the high ZCR area in which, if the ZCR index is greater than the ZCR threshold, then the ZCR index is moved to the first point. Lastly it transforms a sample point-based index for the beginning and ending index.

Figure 5 shows the waveform, zero crossing rate and energy for continuous digit recorded from a male speaker. Also as shown in Fig. 5, the voiced speech can be distinguished from unvoiced speech as it has much greater amplitude displacement when the speech is viewed as a waveform. It also shows a boundary line for begin and end point for each segment.

This endpoint technique managed to show the voiced speech and unvoiced speech (including silence). Furthermore this endpoint detection algorithm has been tested at various places [12] and also tested on Malay digits [13] showing good segmentation for male and female speakers with a reasonable accuracy rate of 87.5%.

For labeling the segmented speech frame, the ZCR and energy are applied to the frame. Unfortunately it contains some level of background noise due to the fact that energy for breath and surroundings can quite easily be confused with the energy of a fricative sound[14]. For voiced speech, energy is high and the ZCR are low. On the other hand, for unvoiced speech the energy is low and the ZCR are high.

Feature Extraction: In this project Mel Frequency Cepstral Coefficients (MFCC) is chosen because of the sensitivity of the low order cepstral coefficients to overall spectral slope and the sensitivity properties of the high-order cepstral coefficient [15]. Currently it is the most popular feature extraction method [15, 16]. MFCC is produced after the recorded signal is pre-emphasized, framed and Hamming windowed. Then the signal is normalized and lowpass filtered. Lowpass filter is used to remove the potential artificial high frequencies appearing in their modulation spectrum due to transmission errors.

The Hamming window is calculated after getting the results from the endpoint process. The equation used is as follows:

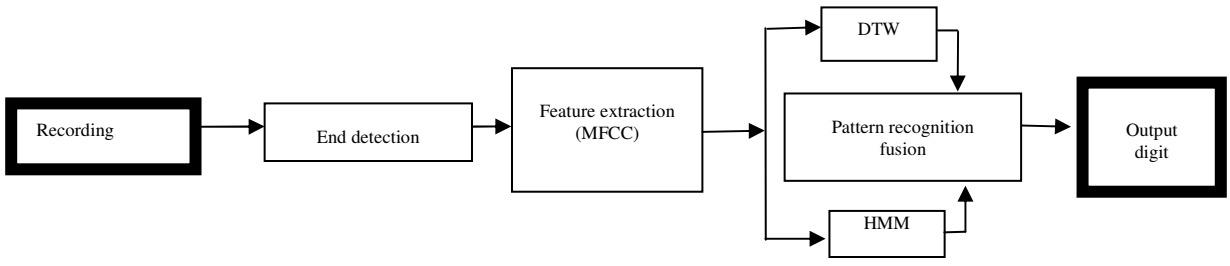

Fig. 3: Block diagram for decision fusion on Malay isolated digit recognition using DTW and HMM

$$w(n) = \frac{\alpha_w - (1-\alpha_w)\cos(2\pi n/(N_s-1))}{\beta_w} \quad (4)$$

where α_w is equal 0.54, meanwhile β_w, functions to normalized the energy through the operation so that the signal will not change. For the purpose of front end to obtain the desired frequency resolution on a Mel scale the simple Fourier Transform (FT) is used. The average spectral magnitude for each amplitude coefficient then is calculated as:

$$S_{avg}(f) = \frac{1}{N}\sum_{n=0}^{N} w_{FB^{(n)}}|S(f)| \quad (5)$$

where the number of samples to get the average value is denoted as N, weighting function is denoted as $w_{FB}^{(n)}$ and magnitude of the frequency computed by the Fourier transform is denoted as $|S(f)|$.

The cepstral coefficient is computed to minimize the non-information bearing variability from that amplitude via the following calculations:

$$c(n) = \frac{1}{N}\sum_{k=0}^{N} \log|S_{avg}(k)|\ e^{j\frac{2\pi}{N}kn}, 0 \le n \le N-1 \quad (6)$$

where the average signal value in the k^{th} is denote as S_{avg}.

Dynamic Time Warping (DTW): DTW is one of the main algorithms in this system for recognition after HMM. Due to the wide variations in speech between different instances of the same speaker, it is necessary to apply some type of non-linear time warping prior to the comparison of two speech instances. DTW is the preferred method for doing this, whereby the principles of dynamic programming can be applied to optimally align the speech signals. On the other hand, for detecting similar shapes with different phases, DTW has been used to calculate more robust distance for time series data. Indeed it can be used to measure similarity between sequences of different lengths. Because of these advantages many researchers use DTW such as for generic analysis and mining tasks on time series data, voice recognition and signature verification [5, 18]. The distance metric used is a Euclidean distance for the cepstral coefficients over all frames after DTW is applied to align the frames optimally. The distance metric between frame *i* of the test word T_{MFCC} and frame j of the reference word R_{MFCC} is calculated as:

$$D_{ij} = \left(\frac{1}{p}\right)\sqrt{\sum_{k=1}^{p}\left(T_{MFCC}(i,k) - R_{MFCC}(j,k)\right)^2} \quad (7)$$

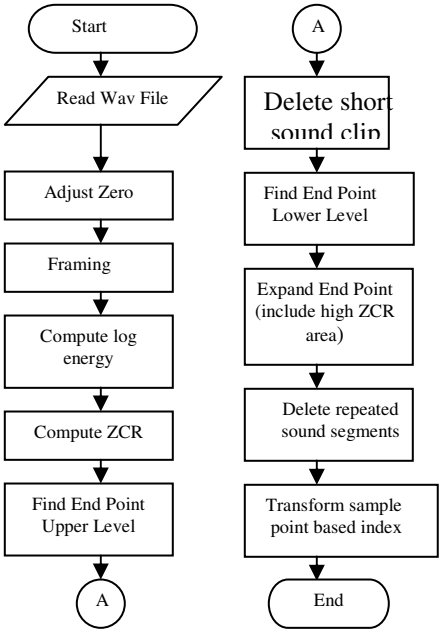

Fig. 4: Flowchart of Segmentation

This DTW algorithm has been tested with 80.5% correctness[19] and one of the screens shot is shown in Fig. 6. But for this fusion system the distance is calculated as:

$$D_{ij} = T_{MFCC}(i,k) - R_{MFCC}(j,k) \quad (8)$$

for the purpose to process one digit only. This distance will be used by decision fusion to process the weight mean vector for one digit.

Hidden Markov Model (HMM): HMM is typically an interconnected group of states that are assumed to emit a new feature vector for each frame according to an emission probability density function associated with that state. Viterbi algorithm is the most suitable for the estimation the parameters for HMM on the maximum likelihood criterion. [14]. In HMM the expression is defined as $\lambda = (A, B, \pi)$. *A* is denoted by a state transition probability matrix, *B* is denoted as output probability matrix and π denoted as initial state probability. The probability of the observation sequence $P(o|\lambda)$ is given multidimensional observation sequences o, known as feature vectors.

For word-level HMM, the recognizer computes and compares all the $P(o|\lambda)$ where (v = 1,2,…,W) and W is the digit word models. For left-to right HMMs, $P(o|\lambda)$ is computed using the Log-Viterbi algorithm as follows[20]:

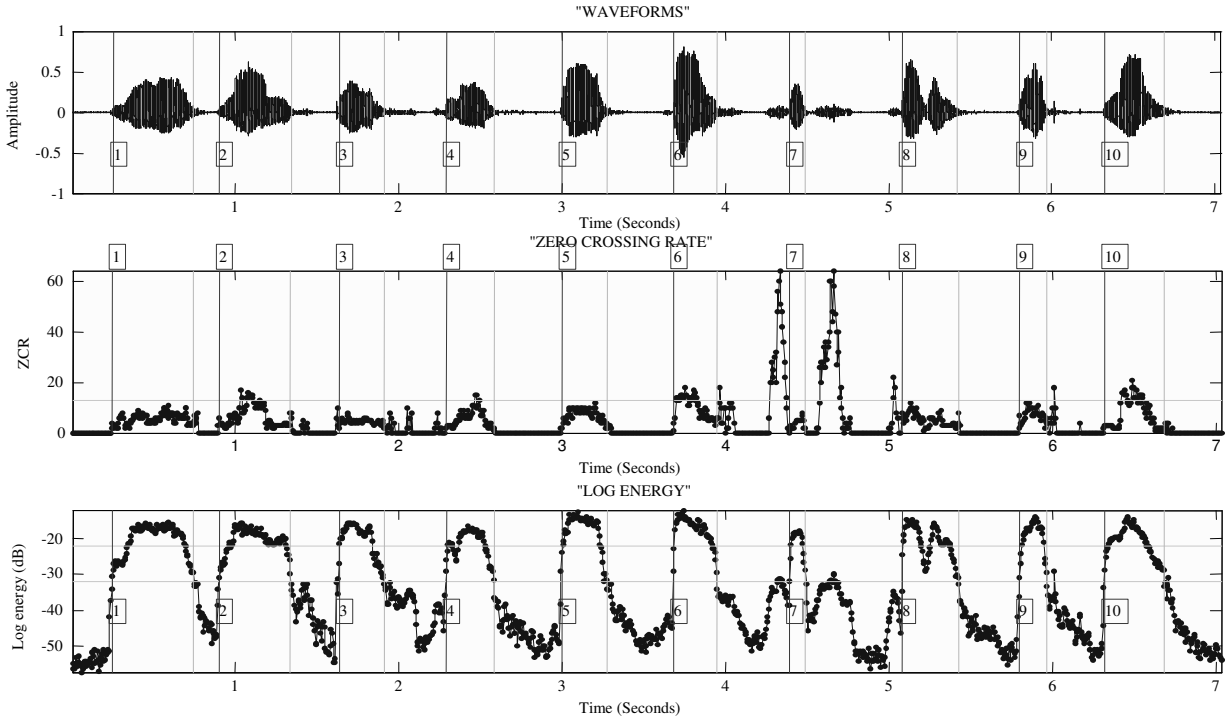

Fig. 5: The waveform, zero crossing rate and energy for continuous digit spoken in wave file recorded from one of the speakers

For initialization,

$$\delta_1(j) = \log \pi_j + \log b_j(o_1) \quad \text{for } t = 1,\ 1 \le j \le N \tag{9}$$

For recursion,

$$\delta_t(j) = \max_{i=j-1,j}\left[\delta_{t-1}(i) + a_{ij}\right] + \log b_j(o_t) \quad \text{for } 2 \le t \le T,\ 1 \le j \le N \tag{10}$$

and for termination,

$$p(o|\lambda_v) = \max_{1 \le i \le N}\left[\delta_T(i)\right] \quad \text{for } t = T \tag{11}$$

The acronym used in the algorithm:

N is number of states,

T is number of frames for feature vectors $o = [o_1, o_2, ..., o_T]$, a_{ij} is state transition between i and j

A = $\{a_{ij}\}$ are their N-by-N matrix,

B = $\{\log b_j(o_t)\}$ is a *N*-by-*T* matrix in log output probability and $\delta_t(j)$ is the likelihood value at the time index t and state j

Pattern Recognition Fusion HMM and DTW: The pattern recognition fusion method used to fuse the results of DTW and HMM is weight mean vector. DTW measures the distance between recorded speech and a template, expanding or shrinking the temporal axis of the target to find the path or warping function which maximizes the similarity between the two speech signals. The distance of the signals is computed at each instant along the warping function. Meanwhile, HMM trained cluster and iteratively moves between clusters based on their likelihoods given by the various HMMs. The weight mean vectors equation used is as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} + \left|w_i - w_{i+1}\right| \tag{12}$$

which expands to,

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_1 + w_2 + ... + w_n} + \left|w_i - w_{i+1}\right| \tag{13}$$

where

w_1 is query recognition rate in HMM test phase,

w_2 is query recognition rate in DTW test phase,

x_n is the real time value of recorded speeches and

$\bar{x}$ is weight mean vector.

For example if recognition percentage for HMM is h and for DTW is d for one digit, then in the fusion model after the query is recognized by DTW and HMM individually, the final percentage is calculated as follows:

$$\bar{x} = (((h * w_1) + (d * w_2)) + |h-d| \quad (14)$$

RESULTS AND DISCUSSION

We have evaluated our algorithm using the data described in the methodology section. The recognition algorithms HMM, DTW and DTW-HMM pattern recognition fusion is then tested for the percentage of accuracy. The test is limited to Malay digits from 0 to 9. Random utterance of digits is done and the accuracy of 100 samples is analyzed. The results obtained from the accuracy test is about 80.5% of accuracy for DTW and 90.7% for HMM and 94% for pattern recognition fusion. The results obtained are shown in Table 1.

Meanwhile for the robustness, we add Gaussian noise to the original speech signals. Table 2. shows the comparison digit recognition percentage after the noise with various signal to noise ratios (SNRs). Amongst the SNR that we have chosen: SNR greater than 30dB (original speech), 20dB, 15dB, 10dB and 5dB. The results show that pattern recognition fusion is better than stand alone recognition even in noisy conditions.

CONCLUSION

This paper has shown a speech recognition algorithm for Malay digits using MFCC vectors to provide an estimate of the vocal tract filter. DTW and HMM are the two recognition algorithms used. DTW is used to detect the nearest recorded voice. Meanwhile HMM is used to emit a new feature vector for each frame according to an emission probability density function associated with that state. The results showed a promising Malay digit speech recognition module. This paper has shown that the fusion technique can be used to fuse the pattern recognition outputs of DTW and HMM. Furthermore it also introduced refinement normalization by using weight mean vector to get better performance with accuracy of 94% on pattern recognition fusion HMM and DTW. Unlikely accuracy for DTW and HMM, which is 80.5% and 90.7% respectively. The percentage of the recognition can be increased by focusing on tweaking the cut-off values used by the algorithm to label the different parts of speech especially on breathy-voice female speakers. This is because the ZCR has a low value for silence and voiced speech, therefore there is more chance of an error between these values, but energy is only high when voiced speech occurs.

Table 1: comparison digit recognition Accuracy Test Result

Word	Accuracy HMM	Accuracy DTW	Accuracy Fusion
KOSONG	97%	65%	92%
SATU	86%	65%	86%
DUA	93%	80%	97%
TIGA	86%	65%	86%
EMPAT	86%	100%	100%
LIMA	87%	75%	92%
ENAM	86%	100%	100%
TUJUH	86%	65%	86%
LAPAN	100%	95%	100%
SEMBILAN	100%	95%	100%
Average	90.7%	80.5%	94%

Table 2: Comparison digit recognition accuracy percentage after insertion of various values Snr

Recognition Alogrithm	SNR >30dB	SNR 20dB	SNR 15dB	SNR 10dB	SNR 5dB
HMM	91	85	73	53	34
DTW	81	73	63	45	28
Fuse	94	87	74	54	34

ACKNOWLEDGEMENTS

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.

REFERENCES

1. Britannica, 2007. Encyclopedia Britannica Online, http://www.britannica.com/eb/ article-9050292.
2. Le, A, 2003. Rich Transcription 2003: Spring speech-to-text transcription evaluation results, Proc. RT03 Workshop, 2003.
3. Le, A, J. Fiscus, J. Garofolo, M. Przybocki, A. Martin, G. Sanders and D. Pallet, 2002. The 2002 NIST RT evaluation speech-to-text results, in Proc. RT02 Workshop, 2002.
4. Sakoe, H., S. Chiba, 1975. Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. ASSP 26: 43-49, Feb. 1978.
4. Zhu, Y. and D. Shasha, 2003. Warping indexes with envelope transforms for query by humming. Proc. of the ACM SIGMOD Int. Conf. on Management of Data. San Diego, California, pp: 181-192.

6. Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition, IEEE transactions Speech Audio Processing, vol. 2, pp: 257-285, 1989.
7. Thian, N.P.H., S. Bengio, J. Korczak, 2002. A Multi-Sample Multi Source Model For Biometric Authentication, IDIAP Research Report 02-14, April 2002.
8. Kim, T.Y. and H. Ko, 2005. Bayesian Fusion of Confidence Measures for Speech Recognition, IEEE Signal Processing Letters, Vol 12, No. 12, December 2005.
9. Rabiner, L.R. and M.R. Sambur, 1975. An Algorithm for Determining the Endpoints of Isolated Utterances, Bell System. Tech. J., Vol. 54, pp: 297-3 15.
10. Rabiner, L.R. and R.W. Schafer, 1978. Digital Processing of Speech Signals, Prentice-Hall Inc.
11. Analog Devices Inc., 1992. Digital Signal Processing Applications using the ADSP-2100 Family Vol. 2, Prentice Hall.
12. Al-Haddad, S.A.R., S.A. Samad and A. Hussain, 2006. Automatic Digit Boundary Segmentation Recognition, M2USIC 2006, Petaling Jaya, Selangor, 16-17 November 2006.
13. Al-Haddad, S.A.R., S.A. Samad, A. Hussain, 2006. Automatic Segmentation for Malay Speech Recognition, Prosiding Seminar Penyelidikan Siswazah 2006, 29-30 Ogos 2006, Bangi, Selangor.
14. Gold, B. and N. Morgan, 2000. Speech and Audio Signal Processing, John Wiley and Sons, USA.
15. Sh-Hussain Salleh, Hong Kai Sze, Tan Tian Swee, 2002. Design and Development of Speech-Control Robotic Manipulator Arm, Seventh Int. Conference on Control, Automation, Robotics And Vision (ICARCV'OZ), Dec 2002, Singapore.
16. Zhu, Q., A. Alwan, 2000. On the use of variable frame rate analysis in speech recognition, Proc. IEEE ICASSP, Turkey, Vol. III, p: 1783-1786, June 2000.
17. ESTI, 2002. Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithm. ETSI standard document - ES 201 108.
18. Chu, S., E. Keogh, D. Hart and M. Pazzani, 2002. Iterative deepening dynamic time warping for time series. In: Proc of SIAM International Conference on Data Mining.
19. Al-Haddad, S.A.R., S.A. Samad and A. Hussain, 2007. Automatic Recognition for Malay Isolated Digits. The 3rd Int. Colloquium on Signal Processing and its Applications (CSPA 2007), Melaka, Malaysia, March 9-11, 2007.
20. Yoshizawa, S., N. Wada, N. Hayasaka, and Y. Miyanaga, 2002. Scalable Architecture for Word HMM-based Speech recognition and VLSI Implementation in Complete System, IEEE Transactions on Circuits and Systems, Vol. 1, No. 11, November 2002.
21. Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey., V. Valtchev and P. Woodland, 2002. The HTK Book for HTK Version 3.2. Cambridge University Engineering Department, UK, 2002.