Speech Disorder Malay Speech Recognition System

S.A.R. AL-HADDAD
Department Computer and Communication System Engineering
University Putra Malaysia
43400 UPM Serdang, Selangor
MALAYSIA
sar@eng.upm.edu.my

Abstract: - Automatic speech recognition systems have the potential to make hard to understand speech more easily recognizable. Designing a system that recognizes impaired speech is more difficult than a system that recognizes normal speech. The Automatic Malay Speech Recognition for Speech Disorder System is able to recognized impaired Malay words spoken by people who suffer from dysarthria, a motor speech disorder resulting from neuron damage, characterized by poor communication. It is developed using techniques used for normal speech recognition but modified to cater for the speech impairment. A feature extraction technique based on the Mel Frequency Cepstrum Coefficient (MFCC) is used along with artificial intelligent algorithms to recognize the speech. In addition, novel pre-processing steps are required to segment the speech prior to recognition taking into account the speech irregularities. The system requires that the user is registered with the system and the system is then trained to accommodate the user speech pattern. The outputs of the system are the visual display of the corrected words uttered or synthesized audio version of the corrected words.

Key-Words: - speech recognition, dynamic time warping, Mel cepstral coefficient

1 Introduction

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 people 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 is a technique aimed at converting a speaker's spoken utterance into a text string. SR is still far from a solved problem. It was 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]. However, these error rates decrease every year, as speech recognition performance has improved quite steadily. Deng and Huang [4] estimated that performance has improved roughly 10 percent a year over the last decade due to a combination of

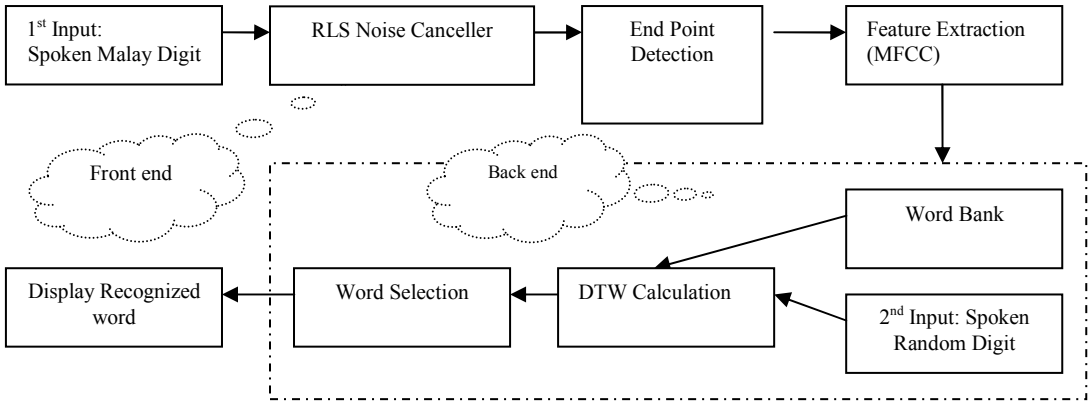

Figure 1. A block diagram of Malay digit recognition

algorithmic improvements and Moore's Law.
Recursive least squares (RLS) algorithm is used to improve the presence of speech in a background of noise. The RLS algorithm provides good performance for models with accurate initial information on a parameter or a state to be estimated [5]. In many applications of noise cancellation, the changes in signal characteristics could be quite fast. This requires the utilization of adaptive algorithms, which converge rapidly. From this perspective the best choice is the RLS [6]. The beginning and end of a word could be detected by the system that processes the word after noise cancellation has been done.
This paper proposes a speech recognition algorithm for Malay digits from 0 to 9. This system consists of speech processing inclusive of digit margin and recognition which uses zero crossing and energy calculation techniques. Mel-Frequency Cepstral Coefficients (MFCC) vectors are used to provide an estimate of the vocal tract filter. Meanwhile dynamic time warping (DTW) is used to detect the nearest recorded voice. This paper is segmented in 4 sections: introduction, methodology, results and discussion, and conclusions.

2 Methodology

The system consists of speech processing and recognition phases as shown in Figure 1. The speech processing phase begins with recording the voice, RLS filtering, endpoint detecting, blocking into frames, frame windowing and calculating MFCC. The MFCC feature 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 [7].

The recognition phase creates a word dictionary or a template of words which is used for the recognition. At the recognition phase the required speech is recorded and processed to detect the speech period and to reduce noise, where the spoken speech is processed while making the word template. The words used in this experiment are Malay isolated digits from 0 to 9 spoken as "KOSONG", "SATU", "DUA", "TIGA", "EMPAT", "LIMA", "ENAM", "TUJUH", "LAPAN" and "SEMBILAN".

RLS method is used in preprocessing stage for noise cancellation as shown in figure 2 [8]. Since the input speech is often corrupted with background interference, a noise canceling stage is inserted before the end point detection process. The operation of the noise canceller can be described by the following set of equations:

$$e = d - y \quad (1)$$

$$d = s + \hat{c} \quad (2)$$

$$y = c_k w_k \quad (3)$$

where :

c is the background noise of any type

$\hat{c}$ is a noise correlated to C

s is a speech signal

d is the desired signal

w_k is the optimum filter weigh matrix

e is the error signal , in the ideal case(clean speech)

k is a time index

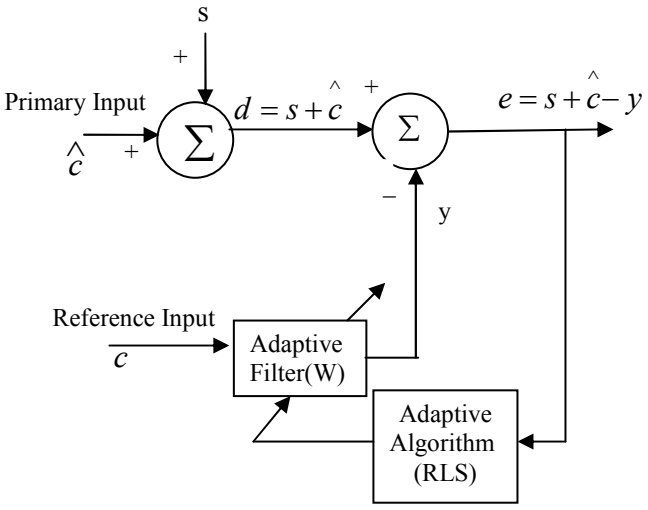

Figure 2: Adaptive noise canceling concept

The heart of the noise canceller is the adaptive filter. The adaptive filter is controlled by RLS

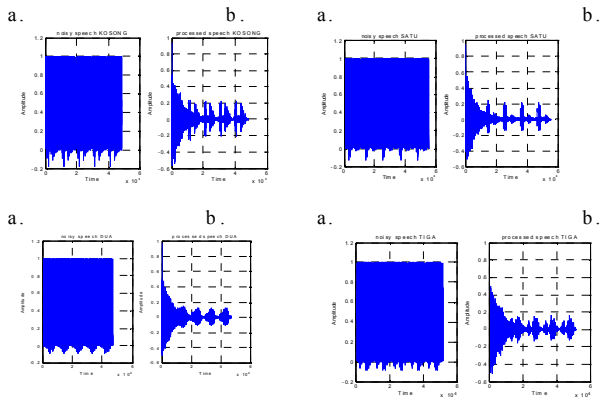

Figure 3: a. Noisy speech ; b. Signal processed by adaptive noise canceller

adaptive algorithm. Due to the non-stationary characteristics of speech signals and background noise, RLS algorithm is used as a controlling algorithm. RLS gives better performance than other algorithms in these circumstances. It can be proven that the weight vector that minimizes the least square criterion is given by Haykin [9] is:

$$\min_{W_n} \sum_{n=0}^{L-1} \| e[n-1] \|^2 = \min_{Wn} \| d_n - X_n^T W_n \|^2 \quad (4)$$

$$\text{where } d_n = \begin{bmatrix} d[n-L+1] \\ \vdots \\ \vdots \\ d[n] \end{bmatrix} \updownarrow L \quad (5)$$

$$\text{and } X_n = \begin{bmatrix} x[n-L+1].....x[n] \\ \vdots \\ x[n-L-L_f+2]....x[n-L_f+1] \end{bmatrix} \updownarrow L_f \quad (6)$$

$$W_n = R_n^{-1} X_n^* d_n \quad (7)$$

k here represents a time index

if $L \geq L_f$. Matrix R_n is defined as $X_n^* X_n^T$. Due to the matrix inverse in equation 7 ,the computational complexity of this least-squares is a function of L_f^3. The complexity can be reduced by recursively updating R_n^{-1} starting from R_{n-1}^{-1} .This leads to the so- called Recursive Least Squares (RLS) algorithm, having a weight update of the form:

$$W_{n+1} = W_n + K_n e[n] \quad (8)$$

in which K_n is the so-called Kalman gain.

Figure 3 shows the results using the RLS adaptive filtering on noisy speech signals. Parts mark a, show the signals of noisy speech and those mark b show the signals after using RLS.

After RLS noise cancellation, zero crossings are calculated, which is 10% of the maximum ZCR. Next, log energy allows us to calculate the amount of energy at a specific instance. For a given 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.

Next the program finds the endpoint upper level by searching from the first point until the energy crosses the upper level energy threshold. Then it deletes short sound clips by eliminating sound length that is less than a certain value. After that, it expands the endpoint lower level by reversing the sound index until it reaches the first point's energy which falls below the low level energy threshold. Next, it expands the endpoint 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 indexes.

This endpoint technique managed to show the voiced speech and unvoiced speech (including silence) segments. Furthermore this endpoint detection algorithm has been tested in various kinds of real noise recorded at various places [10] and also tested on Malay digits [11] which give good segmentation for male and female speakers with a reasonable accuracy rate of 87.5%.

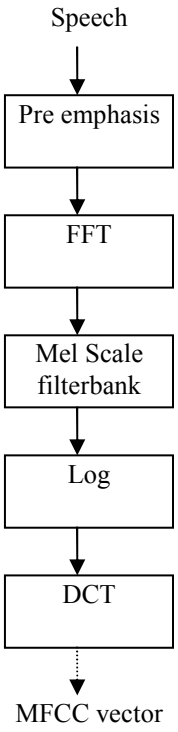

Figure 4. Procedure for calculating MFCC vectors

Before the signal can be made into a template, the signal has to be normalized so that the volume of the speech would not become a factor during speech recognition. The normalization is done by dividing the signal with the maximum absolute value of the signal. The speech signal is then processed in 20 msec (256 point) frames, which were stepped by 10 msec (128 points) between processing frames. Figure 4 shows the stages through which a speech signal passes to be transformed into an MFCC vector which is simplified by Milner and Shao [12]. This conforms to the MFCC standard proposed by the European Telecommunications Standards Institute (ETSI). A few standards for an Automatic Speech Recognition feature extraction are available from the ETSI [13]. The framing method used is Hamming.

After the MFCC process is finished, the results are saved in the sound word bank. Here we used "MalayDigit" as the directory for saving. Then it is tested by using Dynamic Time Warping (DTW). DTW is the main algorithm in this system for recognition. Due to 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 [14].

The application of DTW to isolated digit recognition can be visualized by aligning the processing frames of a reference digit along the y-axis and a test digit along the x-axis. The distance metric is then computed between the frames of the test and reference digit while progressing from the origin at the left bottom corner, up and to the right. The principles of dynamic programming can be applied to find the path, which has the minimum accumulated distance metric. After performing this test using the entire reference vocabulary digit for each test digit, the reference digit with the minimum accumulated distance metric is deemed to be a match. For a speech signal, there are a number of constraints on the search path which can be applied to decrease the complexity of the search.

The primary constraint is that the search should be monotonic, meaning that the path chosen cannot be in negative y or x direction and can also increase only one step at a time. The distance metric is formed by using Euclidean distance for the cepstral coefficients over all the frames, after DTW is applied to align the frames optimally. All paths are given a transition cost of 1. The distance metric between frames i of the test digit T and frame j of the reference digit R is calculated as follows:

$$d(x,y) = \sqrt{\sum_j \left(x_j - y_j\right)^2} \qquad (9)$$

Another constraint, which is the global constraint, is used to restrict the extent of compression or expansion of speech signals over long ranges of

time. The variation of the speech rate of a speaker is considered to be limited in a reasonable range, which means that it can prune the unreasonable search space, and limit the search to the valid region. In order to get the best recognition, the global constraint has to be set to an optimum level; however this is not always possible except with experimentation. To further improve recognition accuracy, the starting point of the search need not be set at the point of origin but at a minimum value of the set predefined margin.

3. Results and Discussions

The system requires users to record numbers 0 until 9 in the Malay language. After that the system saves the recorded voice into a Malay digit directory. Then, the user is required to record any single number between 0 and 9. The input word is then recognized as the word corresponding to the template with the lowest matching score. As an example, in Figure 5 it shows that the word number 2 ("DUA") has the lowest matching score.

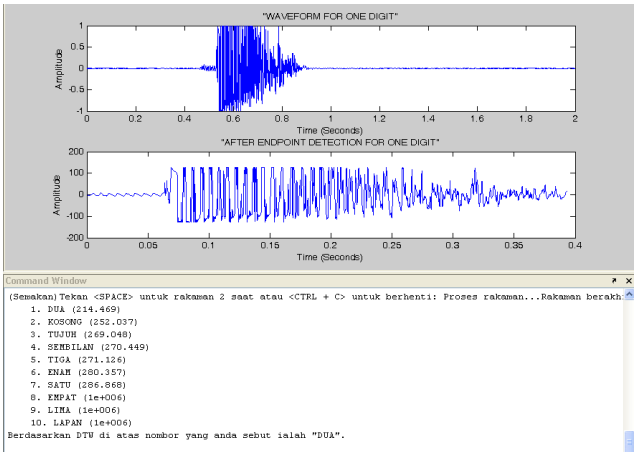

Figure 5. Screenshot of output after recording

The recognition is implemented using DTW where the distance calculation is done between the tested speech and the reference word bank. After the distance is obtained, the path cost is calculated by getting the cheapest path cost with reference to global and local constraints. Recognition accuracy is found to be greatly increased by the implementation of the global constraint and was increased by the application of the local constraint. Figure 6 is a sample of the

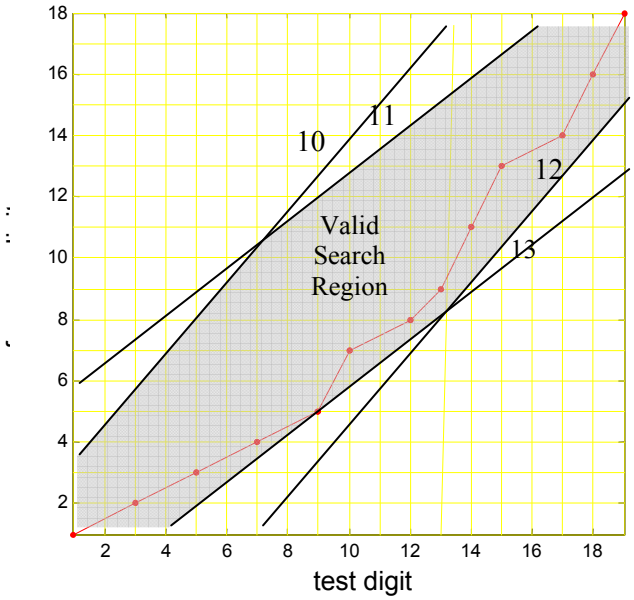

Figure 6. DTW path taken for recognition of utterance of 'DUA'

recognition result of the DTW path cost of the utterance 'DUA'. The shaded area in the figure shows the global constraint (or the valid search area). To increase the accuracy of the recognition, after implementation the appropriate global constraint was found to be as follows:

$$y = 1.61x + 3.84 \quad (10)$$

$$y = 0.93x + 9.07 \quad (11)$$

$$y = 0.86x - 3 \quad (12)$$

$$y = 1.08x - .5 \quad (13)$$

The reason behind setting this global path (from the above equations) is to set a valid search region because the variation of the speech rate of the speaker is considered to be limited in a reasonable range, which means that it can prune the unreasonable search space. The local constraint is as discussed in the methodology section which is monotonicity. Another local constraint applied, which has proven to improve the accuracy, is the start point of the path search.

Table 1. Results of the utterance 'DUA'

Word	Score
DUA	214.469
KOSONG	252.037
TUJUH	269.048
SEMBILAN	270.449
TIGA	271.126
ENAM	280.357

SATU	286.868
EMPAT	1E+006
LIMA	1E+006
LAPAN	1E+006

Table 1 shows the score results of recognition of the utterance 'DUA' which has the corresponding lowest path score and the path taken is as in Figure 6. Table 1 shows that the word 'DUA' has the lowest score of 214.469, which means it is recognized as the word. The closest score to that is the word 'KOSONG'. A limit has to be set in order not to recognize the wrong word. The limit set in this case is 450 paths score because most of the time the result is about 400 paths scores depending on the recording condition of the speech.

Table 2. Accuracy test results without RLS noise canceller

Word	Accuracy
KOSONG	87%
SATU	86%
DUA	100%
TIGA	86%
EMPAT	100%
LIMA	94%
ENAM	100%
TUJUH	88%
LAPAN	100%
SEMBILAN	100%
Average	94.1%

The recognition algorithm is then tested with digits from 0 to 9. Random utterance of numbers is done and the accuracy of 100 samples of numbers is analyzed. The result obtained from the accuracy test is about 80.5%. The results obtained are as displayed in Table 2. Most of the time, the inaccuracy of recognition is due to sudden impulses of noise or a sudden drastic change in the voice tone.

Table 3. Accuracy test results with RLS noise canceller

Word	Accuracy
KOSONG	84%
SATU	84%
DUA	95%
TIGA	84%
EMPAT	100%
LIMA	92%
ENAM	100%
TUJUH	84%
LAPAN	100%
SEMBILAN	100%
Average	92.3%

Next, the RLS noise canceller is used with the algorithm and the results are shown in Table 3. The accuracy is improved to 94.1%.

4. Conclusion

This paper has shown the accuracy of speech recognition algorithm for Malay digits is increased after using RLS noise canceller. MFCC vectors are used to provide an estimate of the vocal tract filter. Meanwhile, DTW is used to detect the nearest recorded voice with appropriate global constraint to set a valid search region because the variation of the speech rate of the speaker is considered to be limited in a reasonable range, which means that it can prune the unreasonable search space. The results showed a promising Malay digit speech recognition module. Recognition with about 80.5% accuracy can be achieved using this method without the RLS noise canceller. The accuracy is increased to 94.1% after using the RLS noise canceller.

References

[1] Britannica, Encyclopedia Britannica Online. (2007). http://www.britannica.com/eb/article-9050292. Accessed date: Aug 8, 2007.

[2] Le, A. (2003). Rich Transcription 2003: Spring speech-to-text transcription evaluation results, Proc. RT03 Workshop, 2003. May 19-20, 2003, Boston, MA, USA. Available from: http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/rt03s-stt-results-v9.pdf. Accessed date: Oct 25, 2007.

[3] Le, A., Fiscus, J., Garofolo, J., Przybocki, M., Martin, A., Sanders, G., and

Pallet, D., "The 2002 NIST RT evaluation speech-to-text results", in Proc. RT02 Workshop, May 7-8, 2002, Vienna, Va, USA. Available from: http://www.nist.gov/speech/tests/rt/rt2002/presentations/rt02_stt_results_v5.pdf. Accessed date: Oct 25, 2007.

[4] Deng, L., and Huang, X. (2004). Challenges in adopting speech recognition. Communications of the ACM, 47(1):69-75.

[5] Park, J.H., Quan, Z.H., Han, S., Kwon, W.H., 2008. New Recursive Least Squares Algorithms without Using the Initial Information, IEICE TRANS. COMMUN., VOL.E91–B, NO.3 MARCH 2008.

[6] Vijayakumar, V.R., Vanathi, P.T., 2007. Modified Adaptive Filtering Algorithm for Noise Cancellation in Speech Signals, Electronics and Electrical engineering, Kaunas Technologija, No 2(74), 2007.

[7] Sheikh, H.S., Hong, K.S., Tan, T.S. (2002). Design and Development of Speech-Control Robotic Manipulator Arm. Proceedings of 7th International Conference on Control Automation, Robotics And Vision (ICARCV 2002), December 2-5, 2002, Singapore, p. 459-463.

[8] Ifeachor, E.C. Jervis, B.W., 2004. Digital Signal Processing – A practical approach, Pearson education, Delhi, 2004.

[9] Haykin, S. 2001. Adaptive Filter Theory Prentice Hall, 4th Ed. 2001

[10] Al-Haddad, S.A.R., Samad, S.A., and Hussain, A. (2006). Automatic digit boundary segmentation recognition. MMU International Symposium on Information and Communication Technology (M2USIC) 2006; Nov 16-17, 2006; Petaling Jaya, Selangor, Malaysia, p. 280-283.

[11] Al-Haddad, S.A.R., Samad, S.A., and Hussain, A. (2007). Automatic Recognition for Malay Isolated Digits", The 3rd International Colloquium on Signal Processing and its Applications (CSPA 2007), Melaka, Malaysia, March 9-11, 2007. CD ROM

[12] Milner, B.P., and Shao, X. (2002). Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model, 7th International Conference on Spoken Language Processing (ICSLP) 2002, September 16-20, 2002, Denver, Colorado, USA, p2421-2424.

[13] European Telecommunications Standards Institute (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, Sophia Antipolis, France.

[14] Gold, B., and Morgan, N. (2000). Speech and Audio Signal Processing. 1st ed. John Wiley and Sons, New York, USA, 537p.