DIGIT BOUNDARY AND RECOGNITION FOR MALAY ISOLATED DIGITS

Syed Abdul Rahman Al-Haddad*, Salina Abdul Samad, Aini Hussain and Khairul Anuar Ishak

Received: Jul 26, 2007; Revised: Nov 1, 2007; Accepted: Nov 5, 2007

Abstract

This paper proposes a speech recognition algorithm for Malay digits from 0 to 9. This system consists of speech processing inclusive of digit boundary and recognition which uses zero crossing and energy 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 with appropriate global constraint 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 algorithm is tested on speech samples that are recorded as a part of a Malay corpus. The results show that the algorithm managed to recognize almost 90.5% of the Malay digits for all recorded words.

Keywords: HMM, DTW, zero crossing technique, log energy, word bounder

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 (Britannica, 2007).

Speech recognition (SR) 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 (Le, 2003). Meanwhile error rates on conversational meeting speech are about 50% higher, and much more under noisy conditions (Le *et al.*, 2002).

However, these error rates go down every year, as speech recognition performance has improved quite steadily. Deng and Huang (2004) estimated that performance has improved roughly 10 percent a year over the last decade due to a combination of algorithmic improvements and Moore's Law.

This paper proposes a speech recognition algorithm for Malay digits from 0 to 9. This system consists of speech processing inclusive

Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering, University Kebangsaan, Malaysia, Bangi, Selangor, Malaysia. E-mail: sar@vlsi.eng.ukm.my

* *Corresponding author*

of digit boundary and recognition which uses zero crossing and energy 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, material and method, results and discussion, and conclusions.

Material and Method

The system consists of speech processing and recognition phases as shown in Figure 1. The speech processing phase begins with recording the voice, endpoint detecting, blocking into frames, frame windowing and MFCC. 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. (Sheikh *et al.*, 2002)

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”.

Next, the processed word is calculated for the cost against the reference template. The speech waveform range is from 300 Hz to 3 kHz, therefore the recording is set to mono, 8 kHz which is sufficient for recognition purposes due to the fact that the maximum frequency of voice is about 4 kHz. Hence, the sampling rate satisfies the Nyquist requirement.

For endpoint detection, the two basic parameters are the Zero Crossing Rate (ZCR) and short time energy. The energy parameter has been used in endpoint detection since the 1970’s (Rabiner and Sambur, 1975). By combining with the ZCR, the speech detection process can be made very accurate (Rabiner and Schafer, 1978). The beginning and ending for each utterance can be detected. A traditional endpoint detection scheme is shown in Figure 2 (Analog, 1992).

The measurements of the short time energy can be defined as follows (Rabiner and Schafer., 1978):

a) logarithm energy:

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

b) sum of square energy:

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

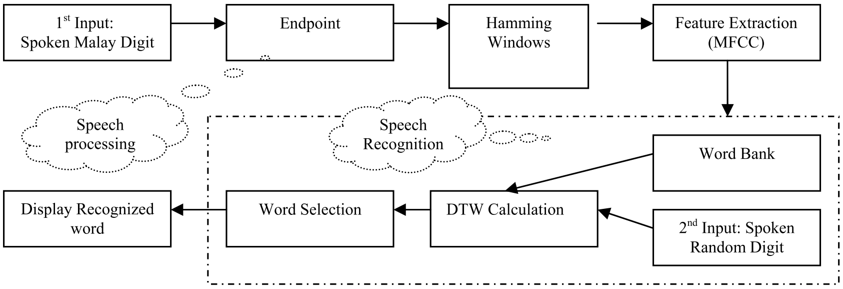


Figure 1. A block diagram of Malay digit recognition

c) sum of absolute energy:

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

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

The flowchart of the endpoint detection is shown on Figure 3. 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 sec pause between each number.

The WAV file is played to hear the sound before it is processed. Then the ZCR is adjusted

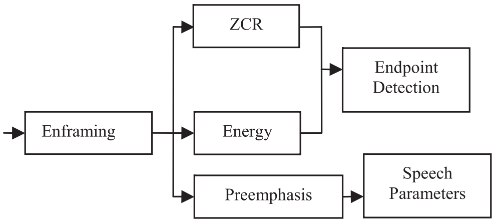


Figure 2. Traditional endpoint detection scheme

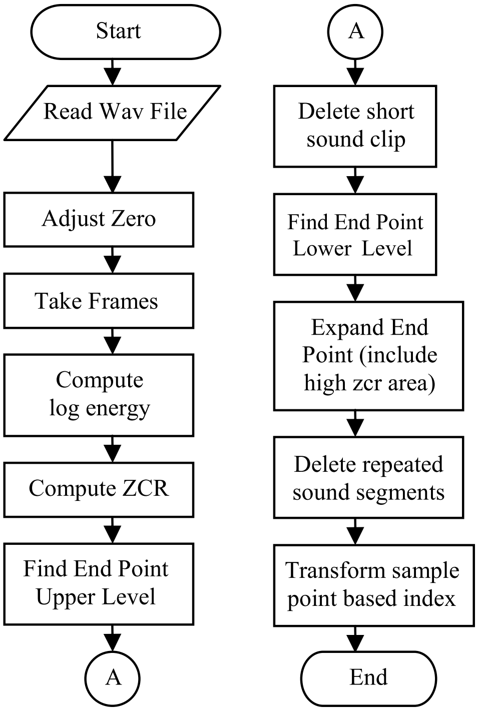

Start
Read Wav File
Adjust Zero
Take Frames
Compute log energy
Compute ZCR
Find End Point Upper Level
A
A
Delete short sound clip
Find End Point Lower Level
Expand End Point (include high zcr area)
Delete repeated sound segments
Transform sample point based index
End

Figure 3. Flowchart of segmentation

to the number of times in a sound sample the amplitude of the sound wave changes the sign by getting their mean (y = y-mean(y)). A tolerance for threshold is included in the function that calculates zero crossings 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 index.

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 (Al-Haddad *et al.*, 2006a) and also tested on Malay digits (Al-Haddad *et al.*, 2006b) which give good segmentation for male and female speakers with a reasonable accuracy rate of 87.5%. For voiced speech, the energy is high and the zero crossing rates are low. On the other hand, for unvoiced speech the energy is low and the zero crossing rates are high.

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 (Gold and Morgan, 2000). Figure 4 shows the waveform, ZCR and energy for continuous digits recorded from a male speaker. Also as shown in Figure 4, 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 the beginning and endpoint for each segment.

As a result, this algorithm performs an almost perfect segmentation for voices recorded by male speakers. For recordings done at noisy places, segmentation problems happen because in some cases the functions produce different values caused by background noise. This causes the cutoff for silence to be raised as it may not be quite zero due to noise being interpreted as speech by the functions. On the other hand under clean speech both the ZCR and short term energy should be zero for silent regions.

Furthermore the way people talk, the volume, and speed also cause problems in detecting the endpoints. 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.

It is necessary then to filter the signal from unwanted noise. 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 in the 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 5 shows the stages through which a speech signal passes to be transformed into an MFCC vector which is simplified by Milner and Shao (2002). 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 (ETSI, 2002). The framing method used is Hamming.

After the MFCC process is finished, the

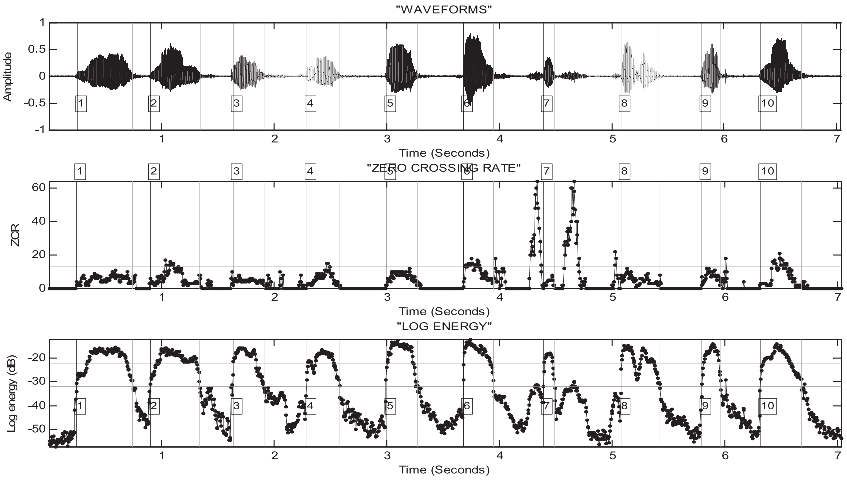


Figure 4. The waveform, zero crossing rate and energy for continuous digit spoken in WAV file recorded from one of the speakers

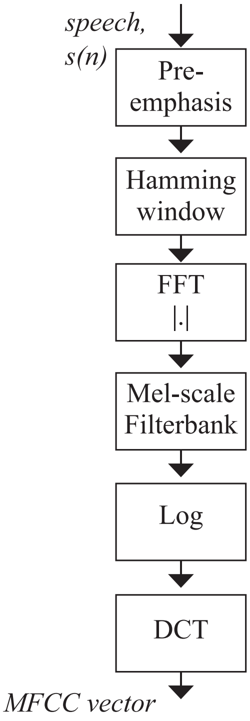


Figure 5. Procedure for calculating MFCC vectors

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 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.

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 as shown in Figure 6. 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 were 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 was calculated as follows.

$$d(x, y) = \sqrt{\sum_j \left(x_j - y_j\right)^2} \quad (4)$$

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

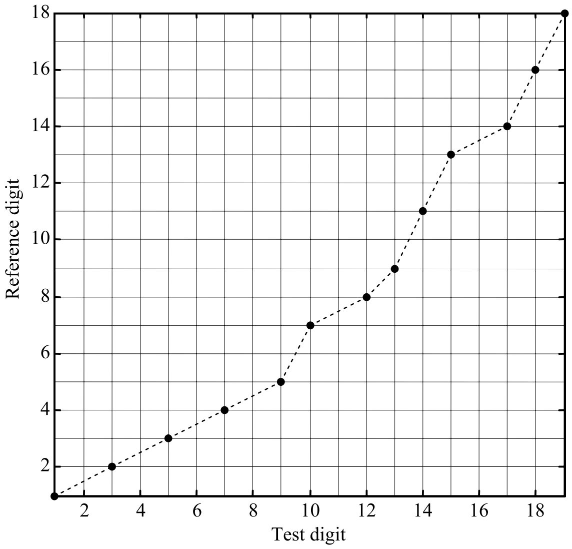


Figure 6. DTW path taken for recognition of utterance of number 2 (spoken in Malay as “DUA”)

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.

Result and Discussion

After running the algorithm, the results are obtained as shown in Figure 7. The system requires the user 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 as shown in Figure 8. The input word is then recognized as the word corresponding to the template with the lowest matching score. As an example, in Figure 8 it shows that the word number 2 ("DUA") has the lowest matching score.

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 9 is a sample of the 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.

Equation:

(1) $y = 1.61x + 3.84$

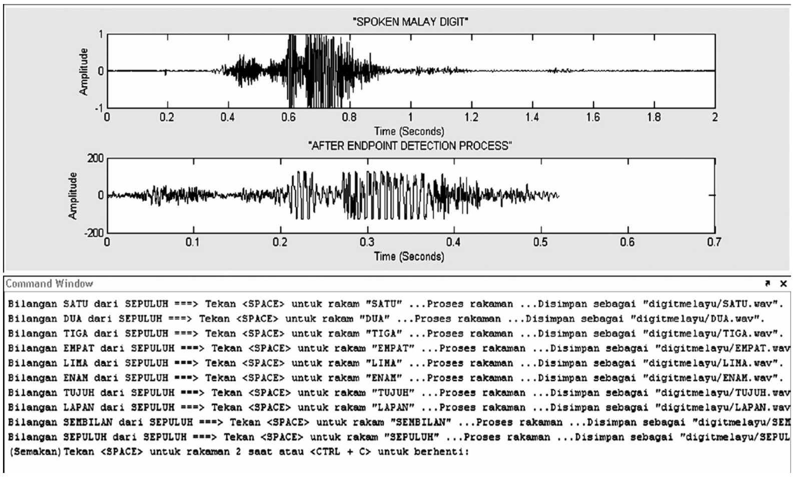
(2) $y = 0.93x + 9.07$

(3) $y = 0.86x - 3$

(4) $y = 1.08x - 0.5$

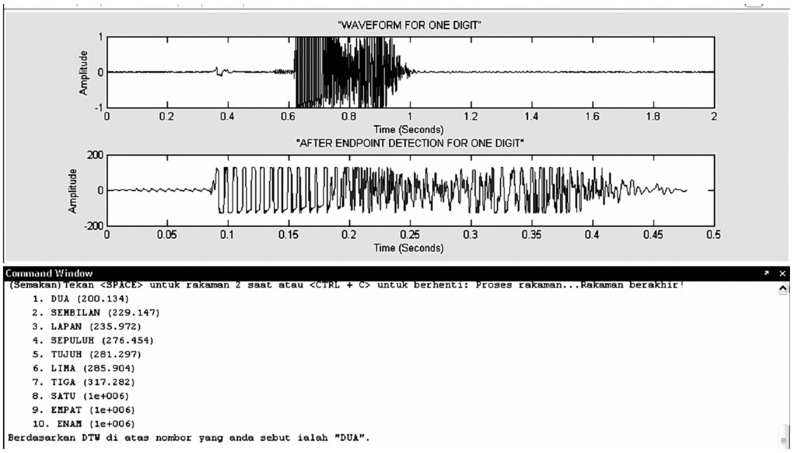
The reason behind setting this global path (from the above equations) is to set a valid search region because the variation of the speech rate of the speaker is considered to be limited in a reasonable range, which means that it can prune the unreasonable search space. The local constraint is as discussed 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 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 8. Table 1 shows that the



```
Bilangan SATU dari SEPULUH ===> Tekan <SPACE> untuk rakam "SATU" ...Proses rakaman ...Disimpan sebagai "digitmelayu/SATU.wav".
Bilangan DUA dari SEPULUH ===> Tekan <SPACE> untuk rakam "DUA" ...Proses rakaman ...Disimpan sebagai "digitmelayu/DUA.wav".
Bilangan TIGA dari SEPULUH ===> Tekan <SPACE> untuk rakam "TIGA" ...Proses rakaman ...Disimpan sebagai "digitmelayu/TIGA.wav".
Bilangan EMPAT dari SEPULUH ===> Tekan <SPACE> untuk rakam "EMPAT" ...Proses rakaman ...Disimpan sebagai "digitmelayu/EMPAT.wav
Bilangan LIMA dari SEPULUH ===> Tekan <SPACE> untuk rakam "LIMA" ...Proses rakaman ...Disimpan sebagai "digitmelayu/LIMA.wav".
Bilangan ENAM dari SEPULUH ===> Tekan <SPACE> untuk rakam "ENAM" ...Proses rakaman ...Disimpan sebagai "digitmelayu/ENAM.wav".
Bilangan TUJUH dari SEPULUH ===> Tekan <SPACE> untuk rakam "TUJUH" ...Proses rakaman ...Disimpan sebagai "digitmelayu/TUJUH.wav
Bilangan LAPAN dari SEPULUH ===> Tekan <SPACE> untuk rakam "LAPAN" ...Proses rakaman ...Disimpan sebagai "digitmelayu/LAPAN.wav
Bilangan SEMBILAN dari SEPULUH ===> Tekan <SPACE> untuk rakam "SEMBILAN" ...Proses rakaman ...Disimpan sebagai "digitmelayu/SEM
Bilangan SEPULUH dari SEPULUH ===> Tekan <SPACE> untuk rakam "SEPULUH" ...Proses rakaman ...Disimpan sebagai "digitmelayu/SEPUL
(Semakan)Tekan <SPACE> untuk rakaman 2 saat atau <CTRL + C> untuk berhenti:
```

Figure 7. Screenshot output after recording ten digits in the Malay language



```
(Semakan)Tekan <SPACE> untuk rakaman 2 saat atau <CTRL + C> untuk berhenti: Proses rakaman...Rakaman berakhir!
    1. DUA (200.134)
    2. SEMBILAN (229.147)
    3. LAPAN (235.972)
    4. SEPULUH (276.454)
    5. TUJUH (281.297)
    6. LIMA (285.904)
    7. TIGA (317.282)
    8. SATU (1e+006)
    9. EMPAT (1e+006)
    10. ENAM (1e+006)
Berdasarkan DTW di atas nombor yang anda sebut ialah "DUA".
```

Figure 8. Screenshot of output after recording digit number 2 (pronounced in Malay as "DUA") where the system will choose the nearest digit inside the storage

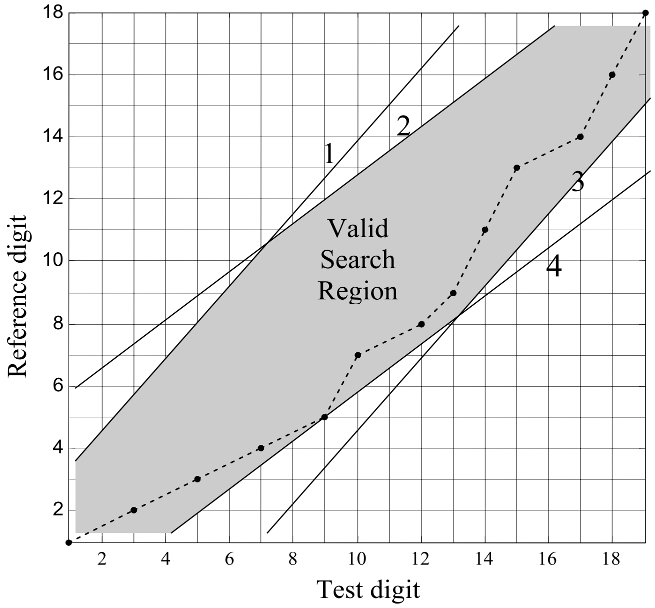


Figure 9. DTW path taken for recognition of utterance of 'DUA'

word 'DUA' has the lowest score of 200.134, which means it is recognized as the word. The closest score to that is the word 'SATU'. 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 score depending on the recording condition of the speech.

The recognition algorithm is then tested for accuracy. The test is limited to digits from 1 to 10. Random utterance of numbers is done and the accuracy of 100 samples of numbers is analyzed. The results obtained from the accuracy test are about 90.5% of accuracy. 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.

Meanwhile for the robustness test, Gaussian noise has been added to the original speech signals. Table 3 shows the comparison digit recognition percentage with various signals to noise ratios (SNR). Amongst the SNR that we have test are 40 dB, 20 dB, 15 dB, 10 dB and 5 dB. The accuracy shows 91%, 84%, 72%, 52.3%, and 33.1% respectively.

Conclusions

This paper has shown a speech recognition algorithm for Malay digits using MFCC vectors to provide an estimate of the vocal tract filter.

Table 1. Results of the utterance 'DUA'

Word	Score
DUA	200.134
SEMBILAN	229.147
LAPAN	235.972
SEPULUH	276.454
TUJUH	281.297
LIMA	285.904
TIGA	317.282
SATU	10^6
EMPAT	10^6
ENAM	10^6

Table 2. Accuracy test results

Word	Accuracy (%)
SATU	80.0
DUA	95.0
TIGA	80.0
EMPAT	100.0
LIMA	90.0
ENAM	100.0
TUJUH	80.0
LAPAN	100.0
SEMBILAN	100.0
SEPULUH	80.0
Average	90.5

Table 3. Accuracy percentage after insertion of various value of SNR

Recognition Algorithm	Accuracy (%)				
	SNR 40dB	SNR 20dB	SNR 15dB	SNR 10dB	SNR 5dB
DTW	91	85	72	53	32

Meanwhile, DTW is used to detect the nearest recorded voice with appropriate global constraint 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 results showed a promising Malay digit speech recognition module. Recognition with about 90.5% accuracy can be achieved using this method, which can be further increased with further research and development 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.

References

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

Al-Haddad, S.A.R., Samad, S.A., and Hussain, A. (2006b). Automatic segmentation for Malay speech recognition. Proceedings of Postgraduate Research Seminar 2006; Aug 29-30, 2006, Bangi, Selangor, Malaysia, p. 235-239.

Analog Devices Inc. (1992). Digital Signal Processing Applications Using the ADSP-2100 Family, Vol. 2. Prentice-Hall Inc., Englewood Cliffs, NJ, USA, 608p.

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

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

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.

Gold, B. and Morgan, N. (2000). Speech and Audio Signal Processing. 1st ed. John Wiley and Sons, NY, USA, 537p.

Kim, D.S., Lee, S.Y., and Kil, R.M. (1999). Auditory processing of speech signals for robust speech recognition in real world noisy environments. IEEE Trans. Speech and Audio Proc., 7(1):55-69.

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.

Le, A., Fiscus, J., Garofolo, J., Przybocki, M., Martin, A., Sanders, G., and Pallet, D. (2007). The 2002 NIST RT evaluation speech-to-text results. 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.

Milner, B.P. and Shao, X. (2002). Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. Proceedings of the7th International Conference on Spoken Language Processing (ICSLP) 2002; Sept 16-20, 2002, Denver, Colorado, USA, p. 2,421-2,424.

Rabiner, L.R. and Sambur, M.R. (1975). An algorithm for determining the endpoints of isolated utterances. The Bell System Technical J., 54(2):297-315.

Rabiner, L.R. and Schafer, R.W. (1978). Digital Processing of Speech Signals. 1st ed. Prentice-Hall Inc., Englewood Cliffs, NJ, USA, 509p.

Sheikh, H.S., Hong, K.S., and Tan, T.S. (2002). Design and development of speech-control robotic manipulator arm. Proceedings of the 7th International Conference on Control Automation, Robotics And Vision (ICARCV 2002; Dec 2-5, 2002; Singapore, p. 459-463.