Jurnal Teknologi, 35(D) Dis. 2001: 65–76

SPEAKER-INDEPENDENT MALAY SYLLABLE RECOGNITION USING SINGULAR AND MODULAR NEURAL NETWORKS

TING HUA NONG[1], JASMY YUNUS[2] & SHEIKH HUSSAIN SHAIKH SALLEH[3]

Abstract. The paper investigates the use of Singular and Modular Neural Networks in classifying the Malay syllable sounds in a speaker-independent manner. The syllable sounds are initialized with plosives and followed by vowels. The speech tokens are sampled at 16 kHz with 16-bit resolution. Linear Predictive Coding (LPC) is used to extract the speech features. The Neural Networks utilize standard three-layer Multi-Layer Perceptron (MLP) as the speech sound classifier. The MLPs are trained with stochastic Back-Propagation (BP). The weights of the networks are updated after presentation of each training token and the sequence of the epoch is randomized after every epoch. The speech training and test tokens are obtained from 25 (17 females and 8 males) and 4 (all females) Malay adult speakers respectively. The total training and test token number are 1600 and 320 respectively. The result shows that modular neural networks outperform singular neural network with a recognition rate of about 92%.

Keywords: Back-propagation, Linear Predictive Coding, Malay syllables, modular Neural Networks, Multi-layer Perceptron, singular Neural Network, speaker-independent

Abstrak. Kertas kerja ini mengkaji penggunaan Rangkaian Neural tunggal dan bermodul dalam mengelaskan bunyi silabus Melayu secara penutur-bebas. Bunyi-bunyi silabus ini dikawalkan dengan plosif dan diikuti oleh vokal. Bunyi-bunyi ini disampelkan pada kadar 16 kHz dengan resolusi sebanyak 16 bit. Pengekodan Ramalan Linear ataupun LPC digunakan untuk memperoleh ciri-ciri bunyi. Rangkaian-rangkaian Neural tersebut menggunakan Perceptron multi-aras (MLP) yang berasas tiga sebagai pengelas bunyi. MLP dilatih dengan perambatan-balik (BP) jenis stokastik. Berat-berat rangkaian Neural dikemaskinikan sejurus selepas setiap bunyi disembahkan dan urutan setiap kala diserampangkan. Sampel bunyi untuk latihan dan ujian dikumpulkan daripada 25 (17 perempuan dan 8 lelaki) dan 4 (kesemua perempuan) orang penutur Melayu masing-masing. Jumlah sampel latihan dan ujian ialah 1600 dan 320 masing-masing. Keputusan menunjukkan bahawa rangkaian neural bermodul mengatasi rangkai neural tunggal dengan kadar pengelasan sebanyak 92%.

Kata kunci: Perambatan balik, Pengekodan Ramalan Linear, Silabus Melayu, Rangkaian Neural bermodul, Perceptron multi-aras, Rangkaian Neural tunggal, penutur bebas

1.0 INTRODUCTION

Malay language itself is a unique language. It is one of the agglutinative languages, which allows addition of affixes to the base word to form new words. The affixes can

[1, 2 & 3] Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM, Skudai, Johor, Malaysia

be in the forms of prefixes, suffixes, confixes or infixes. The Malay language is quite different from English language, where in English, the forming of words is due to the changes of phoneme in the base word itself according to its group of words.

Some research have been done in Malay isolated word recognition [5-7] with promising recognition rate. These systems are able to classify Malay digits from zero to nine in speaker-dependent manner. These systems, however, require large vocabulary of words and very huge speech database in order to recognize huge number of Malay words. Thus, they are impractical in continuous speech recognition system.

The isolated word itself is a large form of linguistic unit. The isolated word can be divided into smaller linguistic unit of syllables. A syllable is a structure of a language, which consists of a vowel, 'v' or 'a' vowel with a consonant, 'c' or 'a' vowel with several consonants. There are several syllable structures in Malay language such as CV, VC, CVC, CCVC, CVCC, CCCV and CCCVC. Among the Malay syllables, the structures of Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) are the most common and they can be found almost in every Malay primary word. Since most of the Malay primary words are comprised of syllable of CV and CVC, we propose a Malay syllable recognition system, which is able to classify some Malay syllables of CV structure in a speaker-independent manner.

With its emerging technology in speech recognition, Neural Networks are able to compete with other conventional approaches such as Dynamic Time Warping and Hidden Markov Models [5, 7, 9–11 and 13]. Neural Networks can offer many advantages such as generalization, non-linearity, fault tolerance and ability to learn. Neural networks have been successfully applied in some syllable recognition systems to recognize Japanese and Mandarin syllables [14-15]. Speech recognition is basically a problem of pattern recognition. Since Neural Networks excel in pattern recognition, we are going to investigate the use of the Neural Networks in classifying these Malay syllables. Two network architectures are proposed and their performances are compared.

Section Two explains the Malay speech database development. Speech feature extraction will be discussed in Section Three. Section Four describes the two Neural Network architectures, which are used in the Malay syllable recognition. Performances of the Neural Networks are shown and discussed in Section Five. The discussions are ended with conclusions in Section Six. Lastly, acknowledgements and references are listed in Section Seven and Section Eight, respectively.

2.0 DATABASE DEVELOPMENT

In the experiment, 16 Malay syllables of CV structures are selected and these syllables are initialized with plosive and followed by vowel. In Malay language, the

plosives are comprised of six consonants such as [b], [d], [g], [p], [t] and [k]. Consonants such as [b], [d] and [g] are called the voiced plosives and the other three are called the unvoiced plosives. The voiced plosives differ from the unvoiced plosives in voicing or phonation. Phonation is defined as the vibration of the vocal folds during the production of sound. Voiced plosives have the phonation whereas the phonation is absent in unvoiced plosives.

There are eight vowel sounds in Malay language [5]. The vowel sounds are comprised of four front vowels, one central vowel and three back vowels. The eight vowel sounds are [i], [e], [ʔ], [a], [ʔ], [u], [o] and [ʔ].

The 16 selected Malay syllable sounds are [ba], [bi], [bu], [da], [du], [gi], [ka], [ki], [ko], [ku], [pa], [pi], [pu], [ta], [tʔ] and [ti]. The individual syllable is extracted from its own bi-syllabic word such as "Baba", "Bibi", "Bubu", "Dada", "Dudu" and so on to allow the variations of phonetic context of the sounds.

The speech training tokens are obtained from 25 Malay adult speakers. The speakers consist of 17 females and 8 males. Each speaker contributes two bi-syllabic words for each syllable sound or equivalent to four tokens for each syllable sound. Thus, the total number of the training tokens is 1600, with 100 tokens for each syllable sound. The speaker-independent test tokens are collected from another four Malay adult females. One of the speakers repeats the session a few days later. The total number of test tokens is 320. The total number of speech training and test tokens and their respective number of speakers are shown in Table 1.

Table 1 Training and Test Tokens

Number of	Training Token	Test Token
Each syllable sound	100	20
All syllables	1600	320
Speakers	25	4

The speech tokens are recorded in a normal room environment with ambient noise of 59.60 dB, via a unidirectional microphone. The speech tokens are sampled at 16 kHz with 16-bit resolution. The frequency of interest for voiced sounds such as vowels and voiced plosives is in the region from 0 to 4 kHz. The region from 4 to 8 kHz is considered to be the most important for unvoiced sounds such as unvoiced plosives [10 and 16]. According to Nyquist theorem, in order to capture a signal, the sampling rate must be at least 2 times higher than the frequency of the signal. The sampling rate of 8 kHz is not able to capture this important region of the unvoiced plosive sounds effectively. Thus, in order to capture both voiced and unvoiced plosive sounds, 16 kHz sampling rate is used.

3.0 SPEECH FEATURE EXTRACTION

Linear Predictive Coding (LPC) is used to extract the speech features from the speech tokens. The LPC coefficients are then converted to cepstral coefficients, which are more reliable and robust than the LPC coefficients [8 and 9].

The start point of the speech tokens is determined by comparing their frame energy-power with predefined thresholds. The token duration is taken around the onset of the Consonant-Vowel region as shown in Figure 1. The token duration is set at 120 ms, which is the optimal token duration [12]. In order to analyze the dynamic speech characteristics, the 120 ms token duration is segmented into 20 frames, with analysis frame length of 25 ms and shift of 5 ms [12].

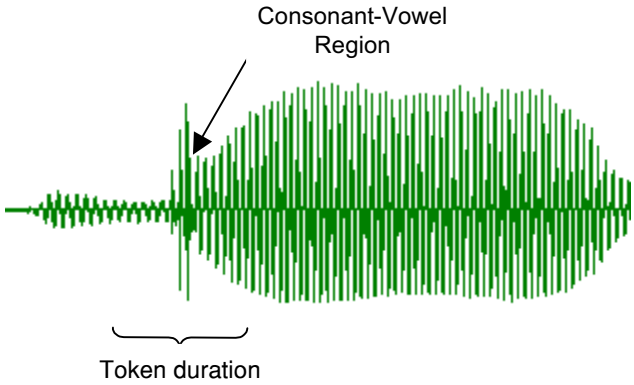

Figure 1 Speech Token Duration

According to Rabiner and Juang [8], a few processes are needed to extract the speech features. First, the speech signals are preemphasized to spectrally flatten the signals. The signals are then Hamming windowed according to the formula:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1 \tag{1}$$

where N is the analysis frame length and $N = 400$ or equivalent to 25 ms at 16 kHz sampling rate. The windowed signals are next autocorrelated according to the formula:

$$r(m) = \sum_{n=0}^{N-1-m} x(n)x(n+m), \qquad m = 0, 1, 2, \ldots, p \tag{2}$$

where p is the order of the LPC analysis and $p = 20$. The zeroth autocorrelation, $r(0)$, is the energy of the analysis frame.

The LPC analysis is performed to convert the autocorrelation coefficients into LPC coefficients. The LPC analysis is implemented using Durbing-Levinson recur-

sive algorithm. The set of equations (3–7) is solved recursively for $i = 1, 2, \ldots, p$, where p is the order of the LPC analysis. The k_i are the reflection or PARCOR coefficients and the a_j are the LPC coefficients.

$$E_0 = r(0) \tag{3}$$

$$k_i = \left[r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(i-j) \right] / E_{i-1}, \qquad 1 \le i \le p \tag{4}$$

$$a_i^i = k_i \tag{5}$$

$$a_j^i = a_j^{k-1} + k_i a_{i-j}^{i-1}, \qquad 1 \le j \le i-1 \tag{6}$$

$$E_i = (1 - k_i^2) E_{i-1} \tag{7}$$

The final solution for the LPC coefficients is given as

$$a_j = a_j^{(p)}, \qquad 1 \le j \le p \tag{8}$$

These LPC coefficients are then converted to cepstral coefficients using the following recursive method [8].

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k}, \qquad 1 \le m \le p \tag{9}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k}, \qquad m > p \tag{10}$$

where m is the order of the cepstral coefficients and $m = 20$. The cepstral coefficients are weighted to reduce their sensitivity to noise.

$$\tilde{c}_m = w_m c_m, \qquad 1 \le m \le p \tag{11}$$

$$w_m = \left[1 + \frac{p}{2} \sin\left(\frac{\pi m}{p} \right) \right], \qquad 1 \le m \le p \tag{12}$$

Lastly, the weighted cepstral coefficients are normalized within -1 and +1 using the formula.

$$w_{\text{normalized}} = 2 \left(\frac{w - w_{\min}}{w_{\max} - w_{\min}} \right) - 1 \tag{13}$$

4.0 NETWORK ARCHITECTURE

Two neural network architectures are proposed to classify the syllable sounds: singular Multi-layer Perceptron network (SMLP) and modular Multi-Layer Perceptron networks (MMLPs). Both networks utilize three-layer MMPLP with one hidden layer as the sound classifier. The networks are trained with stochastic BP, where the weights are updated after presentation of each training token. The Neural Networks are trained to reach a minimum root mean square (rms) error, E_{rms}. The E_{rms} is set to a small value, typically 0.005. The E_{rms} is computed over the entire epoch as shown in the equation below.

$$E_{rms} = \sqrt{\frac{1}{PK}\sum_{i=1}^{P}\sum_{j=1}^{K}\left(T_{ij} - O_{ij}\right)^2} \tag{14}$$

where P = number of training tokens
K = number of output neurons
T_{ij} = target values or desired output and;
O_{ij} = actual output.

The E_{rms} is very descriptive when comparing networks with different output layer size and epoch size. All the neural networks are trained at constant small learning rate to guarantee a true convergence as well as trained at constant large momentum to accelerate the training.

The SMLP has input layer of 400 neuron and output layer of 16 neurons. Each output neuron corresponds to one syllable sound. The hidden neuron number is determined through the experiment. The SMLP is trained with hidden neuron number varied from 100 to 400, in a step of 20 at a training error of 0.005. The result shows that the SMLP with hidden neuron number of 240 obtains the best recognition rate. The architecture of the SMLP is shown in Figure 2.

The performance of the SMLP is further improved by determining the optimal learning rate and momentum of the network. The SMLP is trained with different configurations of learning rate and momentum with hidden neuron number of 240. The learning rate is set at 0.001, 0.01, 0.1 and 0.2. The momentum is varied between 0.60 and 0.90, in a step of 0.10. The result shows that optimal learning rate and momentum of the SMLP is 0.10 and 0.70 respectively.

Modular networks are comprised of two singular neural networks: MMLP I and MMLP II. MMLP I is used to classify the voiced Malay syllable sounds ([ba], [bi], [bu], [da], [du], [gi]) whereas MMLP II is used to classify the rest of the syllable sounds, which are the unvoiced syllable sounds. The architectures of MMLP I and MMLP II are shown in Figures 3a and Figure 3b, respectively.

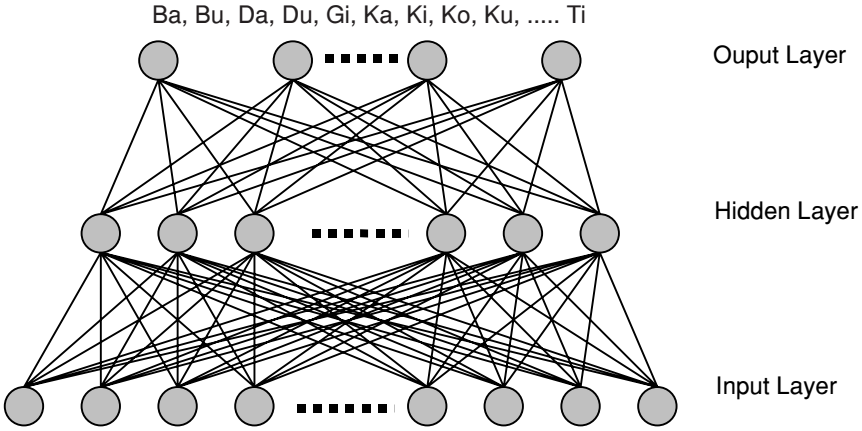

Figure 2 Architecture of the SMLP network

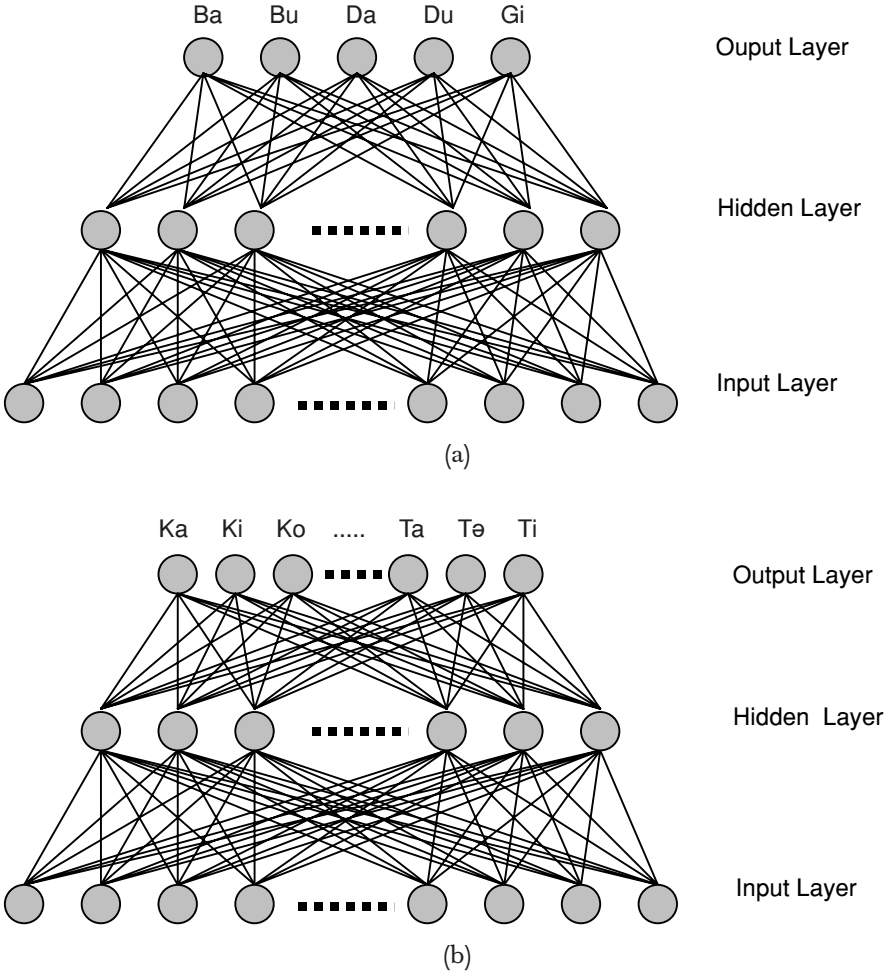

Figure 3 Architecture of (a) MMLP I and (b) MMLP II

MMLP I has an input layer of 400 neuron and an output layer of 5. The MMLP is trained at a training error of 0.005 to search for optimal hidden neuron number, learning rate and momentum. The hidden neuron number is varied from 20 to 260, in a step of 20. The learning rate is set in the range between 0.10 and 0.40, in a step of 0.10. The momentum is varied between 0.60 and 0.90, in a step of 0.10. The results show that the optimal learning rate is 140, with a learning rate of 0.10 and momentum of 0.90.

As for MMLP II, the network has the same size of input and output layer as the MMLP I. Experiments are also conducted to search for the optimal hidden neuron number, learning rate and momentum of the MMLP II. The conditions of the experiments are set as in the previous experiments of MMLP I. The results show that the optimal hidden neuron number of the MMLP II is 160, with optimal learning rate and momentum of 0.10 and 0.80 respectively. The setting of the target values of the MMLPs is shown in Table 2. The activation levels are set at 0.1 and 0.9.

Table 2 Setting of Target Values of MMLPs

Neuron	Ba	Bi	Bu	Da	Du	Gi	Ka	Ki	Ko	Ku	Pa	Pi	Pu	Ta	Tə	Ti
5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.9
4	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.1
3	0.1	0.1	0.1	0.9	0.9	0.9	0.9	0.1	0.1	0.1	0.1	0.9	0.9	0.9	0.9	0.1
2	0.1	0.9	0.9	0.1	0.1	0.9	0.9	0.1	0.1	0.9	0.9	0.1	0.1	0.9	0.9	0.1
1	0.9	0.1	0.9	0.1	0.9	0.1	0.9	0.1	0.9	0.1	0.9	0.1	0.9	0.1	0.9	0.1

5.0 RESULTS AND DISCUSSIONS

The performance of the SMLP is shown in Table 3. The overall performance of the SMLP is 88.75 %. The SMLP is able to fully recognize the syllable sounds of [da], [pa] and [pa]. Among the syllable sounds, [ku] is the worst to be recognized, with a recognition rate of 60.00 %. The result shows that the [ku] tends to be misrecognized as [pu] and [gi] by the SMLP. Besides that, other potential misrecognized pairs are identified as [ka]-[ta], [bu]-[ba] and [ko]-[ku]. The SMLP is better at classifying Malay voiced plosive syllable than the unvoiced plosive syllables. The recognition rate of voiced plosive syllables and unvoiced plosive syllables are 90.83 % and 87.5 %, respectively.

The performances of MMLP I and MMLP II are shown in Tables 4 and 5, respectively. The overall performance of the MMLP I is 90.83 %, with 11 misrecognized tokens. The MMLP I fully recognizes the syllable sounds of [bi] and [da]. The worst recognized sound is [ba], with a recognition rate of 80.00 %.

Table 3 Pattern Classification of SMLP

Syllable	Recognized as																	Acc.,
	Ba	Bi	Bu	Da	Du	GI	Ka	Ki	Ko	Ku	Pa	PI	Pu	Ta	Ta	Ti	Oth	%
Ba	17	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	85.00
Bi	0	19	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	95.00
Bu	3	0	16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	80.00
Da	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	100.00
Du	0	0	0	1	18	0	0	0	0	0	0	0	0	1	0	0	0	90.00
Gi	0	0	0	0	0	19	0	1	0	0	0	0	0	0	0	0	0	95.00
Ka	0	0	0	0	0	0	14	0	0	0	0	0	1	3	1	0	1	70.00
Ki	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	95.00
Ko	0	0	0	0	0	0	0	0	17	3	0	0	0	0	0	0	0	85.00
Ku	0	0	0	0	0	3	0	0	0	12	0	0	5	0	0	0	0	60.00
Pa	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	100.00
Pi	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	1	0	95.00
Pu	0	0	0	0	0	0	0	0	0	0	1	0	19	0	0	0	0	95.00
Ta	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100.00
Tə	0	0	0	0	0	0	0	0	0	0	0	0	1	0	19	0	0	95.00
Ti	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	16	1	80.00

Table 4 Pattern Classification of MMLP I

Syllable	Recognized as							Acc.
	Ba	Bi	Bu	Da	Du	Gi	Others	%
Ba	16	0	3	1	0	0	0	80.00
Bi	0	20	0	0	0	0	0	100.00
Bu	3	0	16	0	1	0	0	80.00
Da	0	0	0	20	0	0	0	100.00
Du	0	0	1	1	18	0	0	90.00
Gi	0	1	0	0	0	19	0	95.00

Table 5 Pattern Classification of MMLP II

Syll-able	Recognized as											Acc.,
	Ka	**Ki**	**Ko**	**Ku**	**Pa**	**Pi**	**Pu**	**Ta**	**T?**	**Ti**	**Oth.**	%
Ka	19	0	0	0	0	0	0	1	0	0	0	95.00
Ki	0	20	0	0	0	0	0	0	0	0	0	100.00
Ko	0	1	17	1	1	0	0	0	0	0	0	85.00
Ku	0	1	3	15	1	0	0	0	0	0	0	75.00
Pa	0	0	0	0	20	0	0	0	0	0	0	100.00
Pi	0	0	0	0	0	19	0	0	0	0	1	95.00
Pu	0	0	0	0	1	1	16	1	1	0	0	80.00
Ta	0	0	0	0	1	0	0	19	0	0	0	95.00
Tə	0	0	0	0	0	0	0	0	20	0	0	100.00
Ti	0	0	0	0	0	1	0	0	0	19	0	95.00

The overall performance of the MMLP II is 92.00 %, with only 16 misrecognized speech tokens. The MMLP II can fully recognize the sounds of [ki], [pa] and [tə]. [ku] again is the worst to be recognized, with a recognition rate of 75.00 %. The [ku] is likely to be misrecognized as [ko]. However, the error is greatly reduced if compared to the SMLPs. The overall performance of the MMLPs is 91.56 %. The performance is 2.81 % higher than the SMLPs. In other words, the MMLPs are able to correctly recognize 9 speech tokens more than the SMLPs.

6.0 CONCLUSIONS

The effectiveness of singular and modular neural networks in classifying Malay syllable sounds in speaker-independent manner has been investigated. The overall performances of the singular and modular neural networks are 88.75 % and 91.56 %, respectively. The study clearly shows that modular neural networks are more effective in classifying the Malay syllable sounds than the singular neural network.

ACKNOWLEDGEMENTS

We would like to thank the Ministry of Science, Technology and Environment of Malaysia for funding this project under the IRPA project no. 09-02-06-0082. Special thanks also to all the speakers involved in the development of the Malay speech database.

REFERENCES

[1] Karim, N. S., F. M. Onn, H. H. Musa, and A. H. Mahmood. 1995. *Tatabahasa Dewan*, New ed. Dewan Bahasa dan Pustaka, Kuala Lumpur.

[2] Hassan, A. 1980. *Linguistik Am Untuk Guru Bahasa Malaysia*. Penerbit Fajar Bakti Sdn. Bhd., Petaling Jaya, Selangor.

[3] Abas, L. 1971. *Linguistik Deskriptif dan Nahu Bahasa Melayu*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

[4] Maris, Y. 1961. *Malay Sound System*. Mimeograph, Kuala Lumpur.

[5] Salleh, S. H. S. 1993. *A Comparative Study of the Traditional Classifier and the Connectionist Model for Speaker Dependent Speech Recognition System*. Unpublished Master Thesis, Universiti Teknologi Malaysia, Johor.

[6] Lim, S. C. 1999. *Isolated Word Speech Recognition Using Hidden Markov Models*. Unpublished Bachelor Thesis, Universiti Teknologi Malaysia, Johor.

[7] Hj Salam, M. S., D. Mohamad and S. H. S. Salleh. 2001. Neural Networks Speaker Dependent Isolated Malay Speech Recognition System: Handcrafted vs Genetic Algorithm, *Proceedings of the 6th International Symposium on Signal Processing and its Applications*. 2: 731–734.

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

[9] Paliwal, K. K. 1990. Neural Net Classifier for Robust Speech Recognition Under Noisy Environments, *International Conference on Acoustics, Speech and Signal Procesing*. ICASSP-90. 1: 429–432.

[10] Lippmann, R. P. 1987. An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine*. 4: 4–22.

[11] Lippmann, R. P. 1989. Review of Neural Networks for Speech Recognition. *Neural Computation*. 1: 1–38.

[12] Ting, H. N., J. Yunus and S. H. S. Salleh. 2001. Malay Syllable Recognition Using Neural Networks. *Proceedings of the International Student Conference on Research and Development*. paper 081 (2001).

[13] Ting, H. N., J. Yunus and S. H. S. Salleh, and E. L. Cheah. 2001. Malay Syllable Recognition Based on Multilayer Perceptron and Dynamic Time Warping. *Proceedings of the 6th International symposium on Signal Processing and Its Applications, ISSPA 2001*. 2: 743–744.

[14] T. Matsuoka, H. Hamada and R. Nakatsu. 2001. Syllable Recognition Using Integrated Neural Networks, *Proceedings of the International Joint Conference on Neural Network, IJCNN 1989*. 1: 251–258.

[15] Zhou L. and S. Imai. 1996. Chinese All Syllables Recognition Using Combination of Multiple Classifiers. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-96*. 6: 3494–3497.

[16] J. Makhoul. 1973. Spectral Analysis of Speech by Linear Prediction. *IEEE Transaction on Audio and Electroacoustics*. AU-21: 140–148.