Malay Speech Recognition using Self-Organizing Map and Multilayer Perceptron

Goh Kia Eng, and Abdul Manan Ahmad
Department of Software Engineering,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia,
81310 Skudai, Johore, Malaysia.
Email: isaac604@hotmail.com, manan@fsksm.utm.my

ABSTRACT

In this paper, a hybrid method based neural network algorithm has been proposed for speech recognition. The proposed method combines Self-Organizing Map (SOM) which known as unsupervised network and Multilayer Perceptron (MLP) which known as supervised network for Malay speech recognition. After the acoustic preprocessing where Linear Prediction Coding (LPC) is used to extract the acoustic information from raw signal, then a 2-dimensional (2D) self-organizing feature map is used as also a feature extractor which acts as a sequential mapping function in order to transform the acoustic vector sequences of speech signal into trajectories. The SOM is used to produce the trajectory vector for classification. The SOM converts the cepstrum vectors into a binary matrix which has the same dimension with the SOM. The idea behind this method is accumulating the all winner node of a syllable utterance in a same dimension map where the winner node is scaled into value "1" and others are scaled into value "0". This result a binary pattern in the 2D map which represent the speech content. The transformation of the feature vector by SOM simplifies the classification task by recognizer using Multilayer Perceptron. The MLP classifies feature vector that each utterance corresponds to. Various experiments were conducted on the 15 Malay syllables by a speaker (speaker dependent system) for conventional technique (MLP only) and the proposed method (SOM and MLP). Our proposed algorithm has achieved better performance where improves the recognition accuracy up to about 4%.

KEYWORDS

Malay Speech Recognition, Neural Network, Self-Organizing Map, and Multilayer Perceptron.

1. Introduction

Neural networks have recently been compared with other pattern recognition classifiers and explored as an alternative to other speech recognition techniques. One of the ideas is to represent the dynamics of a short speech segment in terms of a static model which is a state transition probability matrix derived from the speech segment. In this paper, the static model proposed is employed as an input pattern of multilayer perceptron net [1]. Perceptron attractiveness is due to its well known learning algorithm: Back-propagation [2, 3]. However, there are some difficulties in using perceptron alone. The most major one is that, increasing the number of connections not only increases the training time but also makes it more probable to fall in a poor local minima. It also necessitates more data for training. Perceptron as well as multilayer perceptron usually needs input pattern of fixed length [1]. The reason why the MLP has difficulties is when dealing with temporal information. Since the word has to be recognized as a whole. The word boundaries are often located automatically by endpoint detector and the noise is removed outside of the boundaries. The word patterns have to be also warped using some pre-defined paths in order to obtain fixed length word patterns.

In Kohonen's electronic typewriter [4], Prof. Kohonen uses the clustering and classification characteristics of the SOM to obtain an ordered map from a sequence of feature vectors. The training was divided into two stages, where the first of these was used to obtain the SOM. Speech feature vectors were fed into the SOM until it converged. The second training stage consisted in labeling the SOM, i.e. each neuron of the feature map was assigned a phoneme label. Once the labeling process was completed, the training process ended. Then, unclassified speech was fed into the system, which was then translated it into a sequence of labels. This way, the feature extractor plus the SOM behaved like a transducer, transforming a sequence of speech samples into a sequence of labels. Then, the sequence of labels was processed by some AI scheme in order to obtain words from it.

Usage of an unsupervised learning neural network as well as SOM seems to be wise. Because of its neighboring property, the SOM is found to be suitable. Forming a trajectory and fed to the MLP makes the training and classification simpler and better. This hybrid system consists of two neural based models, a SOM and a MLP. The hybrid system mostly tries to overcome the problem of the temporal variation of utterances.

2. The Design of Speech Recognition

Basically, our speech recognizer is divided into two stages, as shown by the schematic diagram in Figure 1. The Feature Extractor (FE) block shown in this figure generates a sequence of feature vectors, a trajectory in some feature space, which represents the input speech signal. The FE block is the one designed to use the human vocal tract knowledge to compress the information contained by the utterance. The next stage, the Recognizer, performs the trajectory recognition and generates the correct output of syllables. Once the FE block completes its work, its output is classified by the Recognizer module.

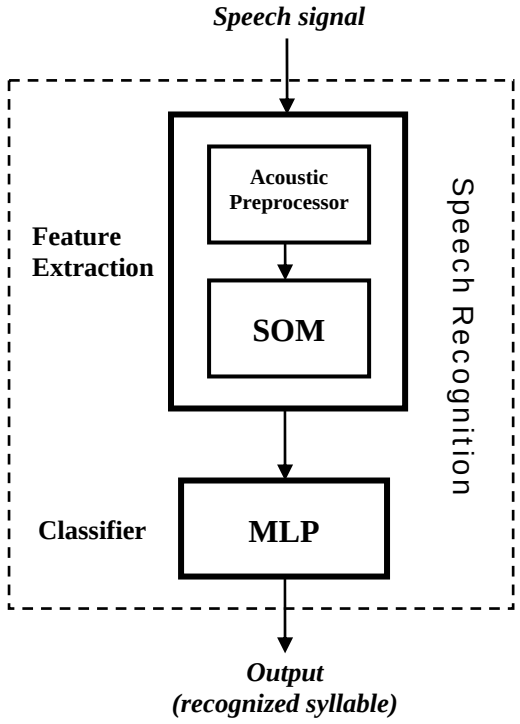

Figure 1: Speech Recognizer Schematic Diagram

2.1 Acoustic Preprocessor

Basically, the FE block works as a transducer that translates the information contained by wave signal into a trajectory in some feature space. First, the incoming wave signal is reduced to a digitized signal through a sampling process. The incoming signal was sampled at 16 kHz with 8 bits of precision. Then, the starting and the ending points of the utterance embedded into the signal are detected using the root mean squared energy (rmse) and zero crossing rate of the signal. The spectrum of the extracted utterance is enhanced by means of a preemphasis filter which boosts the high frequency components. After the signal was sampled, the utterances were isolated, and the spectrum was flattened, each signal was divided into a sequence of data blocks, each block spanning 15ms, or 240 samples, and separated by 5ms, or 80 samples. Next, each block was multiplied by a Hamming window in order to lessen the leakage effects. After that, LPC components are extracted from the filtered blocks. LPC Cepstrum components are then extracted from the LPC vectors.

2.2 Self-Organizing Map (SOM)

T.Kohonen proposed a neural network architecture which can automatically generate self-organization properties during unsupervised learning process, namely, a self-organizing feature map (SOM) [5]. Figure 2 shows the architecture of SOM. All the input vectors of utterances are presented into the network sequentially in time without specifying the desired output. After enough input vectors have been presented, weight vectors from input to output nodes will specify cluster or vector centers that sample the input space such that the point density function of the vector centers tends to approximate the probability density function of the input vectors. In addition, the weight vectors will be organized such that topologically close nodes are sensitive to inputs that are physically similar in Euclidean distance. T.Kohonen has proposed an efficient numerical learning algorithm for practical applications. We used this algorithm in our system.

Denote $M_{ij}(t) = \{ m_{ij}^1(t), m_{ij}^2(t), \ldots, m_{ij}^N(t) \}$ as the weight vector of node (i, j) of the feature map at time instance t; i, j = 1, … , M are the horizontal and vertical indices of the square grid of output nodes, N is the dimension of the input vector. Denote the input vector at time t as X(t), the learning algorithm can be summarized as follows [6]:

1. Initialize all weight vectors to random values in range {-1, +1}.

2. Select the node with minimum Euclidean distance to the input vector X(t)

$$\left\| X(t) - M_{icjc}(t) \right\| = \min_{i,j} \{ \left\| X(t) - M_{ij}(t) \right\| \}. \qquad (1)$$

3. Update weight vectors of those nodes that lie within a nearest neighborhood set of the node (i_c, j_c):

$$M_{ij}(t+1) = M_{ij}(t) + \alpha(t)(X(t) - M_{ij}(t))$$
for $i_c - N_c(t) \le i \le i_c + N_c(t)$ and
$$j_c - N_c(t) \le j \le j_c + N_c(t) \quad (2)$$

$$M_{ij}(t+1) = M_{ij}(t)$$
for all other indices (i, j). (3)

4. Update time t = t + 1, add new input vector and go to (2).

5. Continue until $\alpha(t)$ approach a certain pre-defined value.

In the above equations, $\| \cdot \|$ denotes Euclidean norm, $\alpha(t)$ is a gain term $(0 \le \alpha(t) \le 1)$, $N_c(t)$ is the radius of the neighborhood set around the node (i_c, j_c). The learning constant and neighborhood set both decrease monotonically with time [7]. In our system, we chose $\alpha(t) = 0.001$ and the learning procedure stops when approaches this value. In our experiments, training data include all the frames obtained from the training speech signals and there are over 60000 frames (feature vectors) for each training epoch. The learning algorithm repeatedly presented all the frames until the termination condition is approached. The input vector is the 12 cepstral coefficients as described in Section 2.1. After training, the testing data is fed into the feature map to form a binary matrix. These binary matrixes will be used as input in MLP for classification. The number in a binary matrix determines the number of input node in MLP. A sample of binary matrix is shown in Figure 3.

SOM is used to transform the LPC cepstrum vectors into trajectory in binary matrix form. The LPC cepstrum vectors are fed into a 2-dimensional feature map. The node in the feature map with the closest weight vector gives the response, which is called winner node. The winner node is then scaled into value "1" and other nodes are scaled into value "0". All the winner nodes in feature map are accumulated into a binary matrix with same dimension as the feature map. If a node in the map has been a winner, the corresponding matrix element is unity. Therefore SOM serves as sequential mapping function transforming acoustic vector sequences of speech signal into a two-dimension binary pattern. After mapping all the speech frames of a word, a vector made by cascading the columns of the matrix excites an MLP which has been trained by the backpropagation algorithm for classifying words of the vocabulary. The classification is performed by searching the node giving maximum output value.

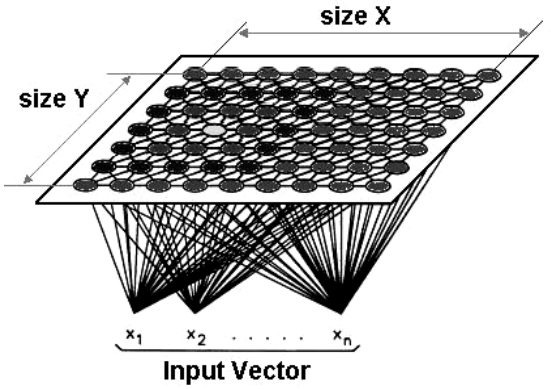

Figure 2: Self-Organizing Map (SOM)

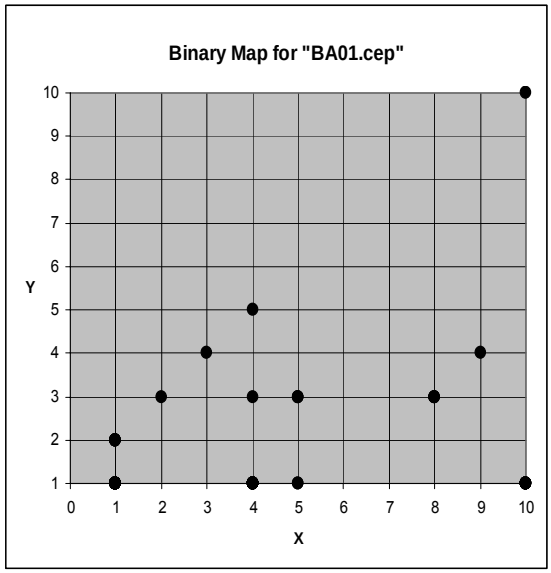

Figure 3: A sample of 10x10 binary matrix where the winner node (●) represent value "1" and others represent value "0".

2.3 Multilayer Perceptron (MLP)

Multilayer Perceptron [8, 9, 10] introduces some hidden units and nonlinear response properties which overcome the limitation of single layer perceptron. In recognition task, input pattern is presented into an internal representation and the output pattern is generated by the internal representation. If there are enough hidden units, MLP can learn to associate pairs of patterns. Once the association has been learnt, the presentation of one member of the pair will produce the other. The connection weights in MLP can be trained by back-propagating errors from the output units.

In our hybrid system, we used two-layer perceptron for testing. The multilayer perceptron has 15 output nodes corresponding to 15 Malay syllables, and 64 hidden nodes which shown in Figure 4. The number of input nodes is same as the dimension of

the input vector (binary matrix). The multilayer perceptron is trained by the backpropagation algorithm with learning rate = 0.2 and momentum coefficient = 0.9. For our training, each syllable has 20 training tokens, thus there are a total of 300 tokens in the training dataset. In all experiments conducted, there are about more than 8000 iterations are needed for the convergence curve of the multilayer perceptron.

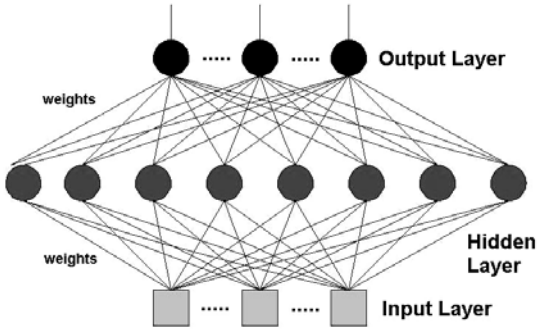

Figure 4: Multilayer Perceptron (MLP)

3. Experimental Results

The experiments were conducted on the 15 chosen Malay common syllables ("BA", "BU", "BI", "KA", "KU", "KI", "MA", "MU", "MI", "SA", "SU", "SI", "TA", "TU", "TI"). The speech was recorded using normal-quality microphone in a laboratory environment with moderate noise-level. The database consists of 600 utterances one speaker (author), where each syllable was uttered for 40 times. 300 of them are used for training and another 300 are used for testing. Comparison between our proposed algorithm (SOM + MLP) and conventional algorithm (Standard MLP) was made by conducting training and testing using different value for parameters. Comparison between SOM using single map and double map was made where double map is made by separating the total frame into 2 parts. There are 3 experiments conducting for different purpose. The details of the experiments are shown in Table 1. The values of the parameter used for network training and testing are shown in Table 2, Table 3 and Table 4.

Table 1: Details of Experiment A, B and C.

Experiments	Objective	Best Result
A	To find out the optimal value for cepstral order	Cepstral order = 20
B	To find out the optimal value for learning rate	Learning rate = 0.2
C	To find out the optimal value for SOM dimension	Dimension = 10 x 10

Table 2: The setting of MLP for Experiment A.

MLP Parameter	
Learning rate	0.25
Input layer	75 frames * cepstral order
Hidden layer	Sqrt(Input * Output)
Output layer	15

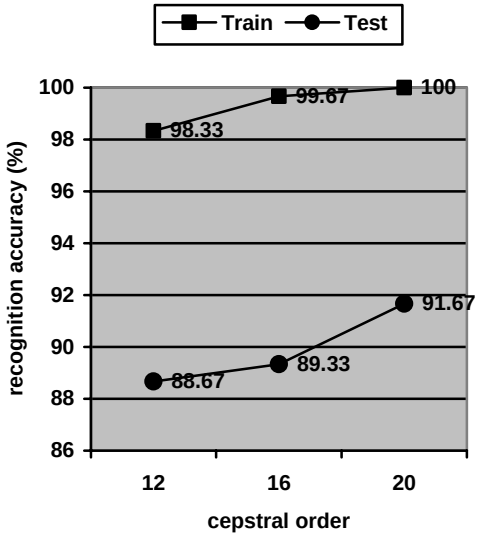

Figure 5: Results for Standard MLP using different cepstral order in Experiment A.

Table 3: The setting of MLP for Experiment B.

MLP Parameter	
Cepstral order	20
Input layer	1500
Hidden layer	150
Output layer	15

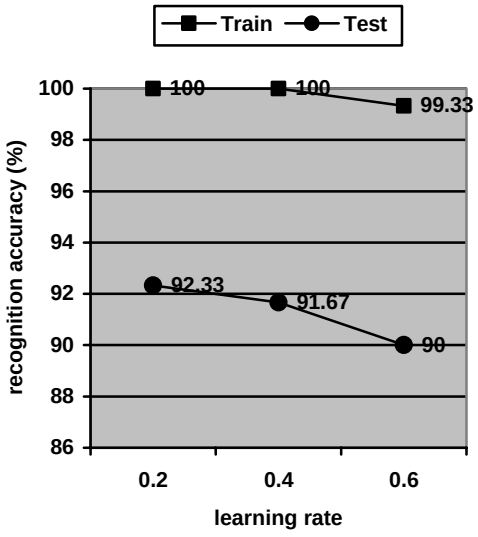

Figure 6: Results for Standard MLP using different learning rate in Experiment B.

Table 4: The setting of SOM + MLP for Experiment C.

Parameter (SOM + MLP)	
Cepstral order	20
Gain (SOM)	0.25
Learning rate	0.2
Input layer	SOM dimension
Hidden layer	Sqrt(Input * Output)
Output layer	15

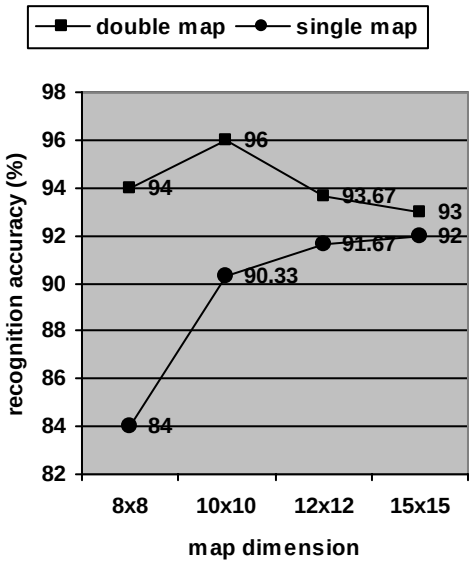

Figure 7: Results for SOM+MLP using different map dimension in Experiment C.

Figure 5-7 shows the graph of the recognition accuracy for Experiment A, B and C. The results show that SOM + MLP algorithm using double map gives the highest accuracy (96%) compared to others. Table 5 shows the accuracy according to each syllable.

Table 5: Results for each syllable in Experiment C which produces the best performance (96%)

Syllable	**Correct**	**Accuracy (%)**
BA	20/20	100
BU	18/20	90
BI	17/20	85
KA	20/20	100
KU	20/20	100
KI	20/20	100
MA	14/20	70
MU	20/20	100
MI	20/20	100
SA	18/20	90
SU	20/20	100
SI	19/20	95
TA	20/20	100
TU	19/20	95
TI	20/20	100

4. Conclusions

In this paper, we have proposed a new way to handle the sequential nature of speech signal using combined self-organizing map and multilayer perceptron for Malay syllables speech recognition. The feature map was trained by Kohonen's self-organization algorithm which simplified the feature vectors by converting them into fixed dimension of binary matrix. The SOM was able to perform good mapping for the MLP in classification task. We compared out system with a conventional system and the experimental results showed that our proposed algorithm improve the recognition accuracy of about 4%. It also showed that our hybrid system being capable for the Malay syllables recognition with overall recognition accuracy of 96%.

5. References

[1] Lippman, R.P. 1989. "Review of Neural Network for Speech Recognition", Neural Computation, Vol. 1, No. 1.

[2] Richard, D., Miall, C. and Mitchison, G. 1989. "The Computing Neuron," Addison-Wesley.

[3] Haykin, S. 1994. "Neural Networks: A Comprehensive Foundation," Macmillan College Publishing Company.

[4] Kohonen, T. 1988. "The 'Neural' Phonetic Typewriter," IEEE Computation Magazine, pp. 11-22. (March).

[5] Kohonen, T. 1984. "Self-Organization and Associative Memory," Springer-Verlag.

[6] Kohonen, T. 1990. "The Self-Organizing Map", Proc. Of the IEEE, Vol. 78, No. 9. (Sept).

[7] Torkkola, K. and Kokkonen, M. 1991. "Using the Topology-Preserving Properties of SOMs in Speech Recognition," Proceedings of the IEEE ICASSP.

[8] Bourland, H. and Wellekens, C. 1987. "Multilayer Perceptron and Automatic Speech Recognition", in Proc. IEEE ICNN (San Diego, CA), IV-407.

[9] Hush, D.R. and Horne, B.G. 1993. "Progress in Supervised Neural Networks", IEEE SP Magazine, (Jan).

[10] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. 1986. "Learning Internal Representations by Error Propagation", in Parallel Distributed Processing, Vol. 1, Chapter 8, Rumelhart, D.E, McClelland, J.L. Eds., Cambridge, MA, MIT Press.