Malay Syllables Speech Recognition Using Hybrid Neural Network

Abdul Manan Ahmad * and Goh Kia Eng**

* Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, Malaysia.
(Tel: +60(07)-5532070; Email:manna@fsksm.utm.my)
** Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, Malaysia.
(Tel: +60(07)-5532070; Email: isaac604@hotmail.com)

Abstract: This paper presents a hybrid neural network system which used a Self-Organizing Map and Multilayer Perceptron for the problem of Malay syllables speech recognition. The novel idea in this system is the usage of a two-dimension Self-organizing feature map as a sequential mapping function which transform the phonetic similarities or acoustic vector sequences of the speech frame into trajectories in a square matrix where elements take on binary values. This property simplifies the classification task. An MLP is then used to classify the trajectories that each syllable in the vocabulary corresponds to. The system performance was evaluated for recognition of 15 Malay common syllables. The overall performance of the recognizer showed to be 91.8%.

Keywords: speech recognition, hybrid neural network, Malay syllables, multilayer perceptron, self-organizing map.

1. INTRODUCTION

Several approaches have been proposed for speech recognition, among them neural networks show to be very effective. The application of neural networks to speech recognition has recently attracted considerable interest. Various neural networks architectures have been largely used for speech recognition task. Many of them are based on perceptron as well as multilayer perceptron [1].

Perceptron attractiveness is due to its well known learning algorithm: Back-propagation [2]. 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 necessitates more data for training [3].

Perceptron as well as multilayer perceptron usually needs input pattern of fixed length. 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.

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 to feed 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. DESCRIPTION OF PROPOSED ALGORITHM

The recognition process of the system is shown in Figure 1. First, input digital speech signals are blocked in frames, acoustic parameters are calculated from each frame of speech by the acoustic preprocessor. This will be explained more detail in Section 3. Then the acoustic vector of the frame is input into the 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 of the 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.

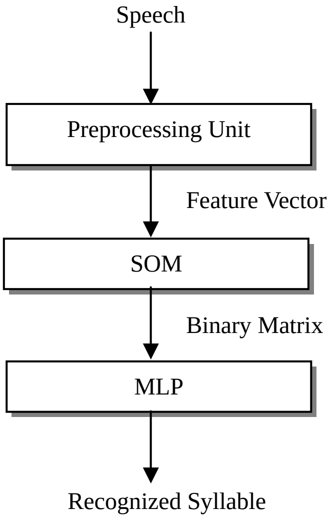

Fig. 1 The architecture of speech recognition system.

3. ACOUSTIC PREPROCESSING

Speech signals of 15 Malay syllables database have been low-pass filtered at 6 kHz and digitized at a 16 kHz sampling rate with 16-bit resolution. Each frame is 320 points (20 ms) and the frame shift is 160 points (10 ms). The preprocessing unit is made up of the following subsections:

- Analog low-pass filter at 6 kHz.
- 16-bit ADC with 16 kHz sampling rate.
- Endpoint Detection using rms energy and zero-crossing rate.
- Pre-emphasis.

- Frame blocking to 20ms per frame with overlap shifting of 10ms.
- Hamming window.
- LPC analysis of order-12.
- Converting LPC coefficients to 12 cepstral coefficients.
- Normalizing and liftering the cepstral coefficients.

4. SELF-ORGANIZING MAP

T.Kohonen [4, 5] proposed a neural network architecture which can automatically generate self-organization properties during unsupervised learning process, namely, a self-organizing feature map (SOM). 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 $M_{ij}(t) = \{ m_{ij}^1(t), m_{ij}^2(t), \cdots, m_{ij}^N(t) \}$ 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:

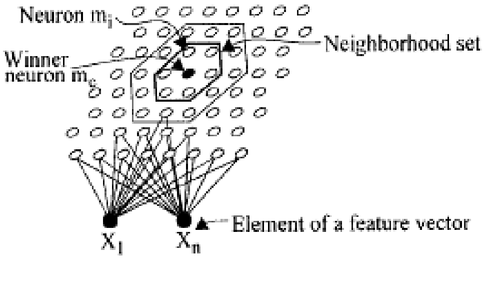

Fig. 2 The architecture of Self-Organizing Map

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)

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

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

$$M_{ij}(t+1) = M_{ij}(t) + \alpha(t)(X(t) - M_{ij}(t)) \quad (2)$$

for $i_c - N_c(t) \le i \le i_c + N_c(t)$ and
$j_c - N_c(t) \le j \le j_c + N_c(t)$

$$M_{ij}(t+1) = M_{ij}(t) \quad (3)$$

for all other indices (i, j).

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 [5]. 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 3.

After training, the testing data is fed into the feature map to form a binary matrix which described before in Section 2. 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.

5. MULTILAYER PERCEPTRON

In our hybrid system, we used two-layer perceptron for testing which is shown in Figure 3. The multilayer perceptron [6] has 15 output nodes corresponding to 15 Malay syllables, and 64 hidden nodes. 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 [2] with learning rate = 0.25 and momentum coefficient = 0.85.

For our training, each word has 20 training tokens, thus there are a total of 200 tokens in the training set. In all experiments conducted, there are about more than 8000 epochs are needed for the convergence curve of the multilayer perceptron. The training set of the MLP, in which every syllable was also presented in equal frequency, was constructed from the binary patterns made by trained feature map.

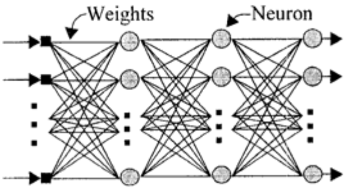

Fig. 3 The architecture of Multilayer Perceptron.

6. EXPERIMENTAL RESULTS

The testing was conducted on the 15 chosen Malay common syllables (“SA”, “SU”, “SI”, “TA”, “TU”, “TI”, “BA”, “BU”, “BI”, “KA”, “KU”, “KI”, “MA”, “MU”, “MI”). The database consists of speech data from 10 speakers (5 male, 5 female), and for each speaker there are 40 repetitions per syllables (20 for training, 20 for testing).

For SOM, we tested different dimension of feature map and tried to find out the optimal dimension for feature map. Each weight vector of SOM was initialized with uniformly distributed random values. There are 20 tokens of each syllables by each speaker used to training the feature map. Then, these training data were converted into binary matrixes which were used to train the MLP by error-backpropagation algorithm. The most common training parameters of SOM and MLP are listed in Table 1.

Table 1
Training parameters of SOM and MLP

Parameter	SOM	MLP
Epoch number	250000	8500
Gain term	0.001	-
Radius	12	-
Learning rate	-	0.25
Momentum	-	0.85
Hidden nodes	-	64

The testing set recognition results are shown in Table 2. Slightly better performance was achieved with a SOM with 12x12 dimension. The usage of SOM in our system has simplified to classification in MLP where the acoustic vector were transformed into fixed-length binary matrix. This also brings a shorter convergence time in training of MLP by backpropagation. The best performance (accuracy) achieved by our hybrid system was 91.8%.

Table 2
Recognition accuracies for training of feature map
in different dimension.

Size of Feature Map	Recognition accuracy
10 x 10	88.2 %
12 x 12	91.8 %
14 x 14	89.3 %
16 x 16	89.1 %
18 x 18	85.9 %
20 x 20	82.6 %

7. CONCLUSIONS

In this paper, we have proposed a hybrid neural network which combining self-organizing feature map and multilayer perceptron for Malay common syllables recognition. The feature map was trained by Kohonen’s self-organization algorithm which simplified the feature vectors by transforming them into fixed dimension binary matrix form. The SOM was able to perform good mapping for the MLP. The results show the hybrid system being capable for the Malay syllables recognition. The hybrid system achieved the overall recognition accuracy of 91.8 %.

8. REFERENCES

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

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

[3] D. R. Hush, B. G. Horne, “Progress in Supervised Neural Networks”, *IEEE SP Magazine*, Jan. 1993.

[4] T. Kohonen, “The ‘Neural’ Phonetic Typewriter,” *IEEE Computation Magazine*, pp. 11-22, March 1988.

[5] T. Kohonen, “The Self-Organizing Map”, *Proc. Of the IEEE*, Vol. 78, No. 9, Sept. 1990.

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