SELF-ORGANIZING MAP AND MULTILAYER PERCEPTRON FOR MALAY SPEECH RECOGNITION

GOH KIA ENG

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

AUGUST 2006

To my beloved mother and father

ACKNOWLEDGEMENT

First of all, I would like to thank my mother and father who have been supporting me and giving me lots of encouragements to complete this thesis. They've been so great and I know there would be no way I could have such a wonderful life without having the love and care from them. Thanks for always been there for me.

A special thank to my supervisor, Prof. Madya Abdul Manan bin Ahmad, for all his guidance and time. Thanks so much for his advices, comments and suggestions on how to improve this research and how to produce a good thesis. He is an understanding and helpful person in helping me to complete this research.

Not forgetting I also would like to take this opportunity to thank all my friends. All the motivations, helps, and supports are fully appreciated. Thanks for being there and listening to my complaints and lend me a helpful hand when I am in troubles.

Last but not least for those who were not mentioned above, I would like you to know that your countless effort and support will always remembered. All credits to everyone! Thank you very much.

ABSTRACT

Various studies have been done in this field of speech recognition using various techniques such as Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Artificial Neural Network (ANN) in order to obtain the best and suitable model for speech recognition system. Every model has its drawbacks and weaknesses. Multilayer Perceptron (MLP) is a popular ANN for pattern recognition especially in speech recognition because of its non-linearity, ability to learn, robustness and ability to generalize. However, MLP has difficulties when dealing with temporal information as it needs input pattern of fixed length. With that in mind, this research focuses on finding a hybrid model/approach which combines Self-Organizing Map (SOM) and Multilayer Perceptron (MLP) to overcome as well as reduce the drawbacks. A hybrid-based neural network model has been developed to speech recognition in Malay language. In the proposed model, a 2D SOM is used as a sequential mapping function in order to transform the acoustic vector sequences of speech signal into binary matrix which performs dimensionality reduction. The idea of the approach is accumulating the winner nodes of an utterance into a binary matrix where the winner node is scaled as value "1" and others as value "0". As a result, a binary matrix is formed which represents the content of an utterance. Then, MLP is used to classify the binary matrix to which each word corresponds to. The conventional model (MLP only) and the proposed model (SOM and MLP) were tested for digit recognition ("*satu*" to "*sembilan*") and word recognition (30 selected Malay words) to find out the recognition accuracy using different values of parameters (cepstral order, dimension of SOM, hidden node number and learning rate). Both of the models were also tested using two types of classification: syllable classification and word classification. Finally, comparison and discussion was made between conventional and proposed model based on their recognition accuracy. The experimental results showed that the proposed model achieved higher accuracy.

ABSTRAK

Banyak penyelidikan telah dijalankan dalam bidang pengecaman suara menggunakan pelbagai teknik seperti *Dynamic Time Warping* (DTW), *Hidden Markov Models* (HMM), *Artificial Neural Network* (ANN) dan sebagainya. Namun demikian, setiap teknik mempunyai kelemahannya masing-masing. Hal ini menyebabkan sistem menjadi kurang tepat. *Multilayer Perceptron* (MLP) merupakan satu rangkaian neural yang terkenal bagi pengecaman suara. Walau bagaimanapun, MLP mempunyai kelemahan di mana akan melemahkan pretasi sistem. Oleh itu, penyelidikan ini menumpu terhadap pembangunan satu model *hybrid* yang menggabungkan dua rangkaian neural iaitu *Self-Organizing Map* (SOM) dan *Multilayer Perceptron* (MLP). Satu model berasaskan rangkaian neural *hybrid* telah dibangunkan bagi sistem pengecaman suara dalam bahasa Melayu. Dalam model ini, SOM yang berdimensi dua digunakan sebagai fungsi pemetaan turutan untuk menukar turutan vector akuastik bagi isyarat suara kepada matrik binari. Hal ini bertujuan untuk mengurangkan dimensi bagi vektor suara. SOM menyimpan nod pemenang bagi suara dalam bentuk matrik di mana nod pemenang diskalakan kepada nilai “1” dan yang lain diskalakan kepada nilai “0”. Hal ini membentukkan satu matrik binari yang mewakili kandungan suara tersebut. Kemudian, MLP mengelaskan matrik binari tersebut kepada kelas masing-masing. Ekperimen dijalankan terhadap model tradisional (MLP) and model hybrid (SOM dan MLP) dalam pengecaman digit (“satu” to “sembilan”) dan pengecaman perkataan 2-suku (30 perkataan yang dipilih). Experimen ini bertujuan untuk mendapat ketepatan pengecaman dengan menggunakan nilai parameter yang berbeza (dimensi cepstral, dimensi SOM, bilangan nod tersembunyi dan kadar pembelajaran). Kedua-dua model ini juga diuji dengan menggunakan dua teknik pengelasan: pengelasan mengikut suku perkataan dan perkataan. Perbandingan dan perbincangan telah dibuat berdasarkan ketepatan pengecaman masing-masing. Keputusan eksperimen menunjukkan bahawa model kami mencapai ketepatan yang lebih tinggi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
ANN	-	Artificial Neural Network
BMU	-	Best Matching Unit
BP	-	Back-Propagation
CO	-	Cepstral Order
CS	-	Conventional System
DR	-	Digit Recognition
DRCS	-	Digit Recognition Conventional System
DRPS	-	Digit Recognition Proposed System
DSOM	-	Dimension of Self-Organizing Map
DTW	-	Dynamic Time Warping
FE	-	Feature Extractor
GPR	-	Geometric Pyramid Rule
HMM	-	Hidden Markov Model
HNN	-	Hidden Node Number
KSOM	-	Kohonen Self-Organization Network
LP	-	Linear Prediction
LPC	-	Linear Predictive Coding
LR	-	Linear Rate
LVQ	-	Learning Vector Quantization
MLP	-	Multilayer Perceptron
PARCOR	-	Partial-Correlation
PC	-	Personal Computer
PS	-	Proposed System
SAMSOM	-	Structure Adaptive Multilayer Self-Organizing Map
SLP	-	Single-layer Perceptron
SOM	-	Self-Organizing Map

TDNN	-	Time-Delay Neural Network
VQ	-	Vector Quantization
WPF	-	Winning Probability Function
WR	-	Word Recognition
WRCS	-	Word Recognition Conventional System
WRCS(S)	-	Word Recognition Conventional System using Syllable Classification
WRCS(W)	-	Word Recognition Conventional System using Word Classification
WRPS	-	Word Recognition Proposed System
WRPS(S)	-	Word Recognition Proposed System using Syllable Classification
WRPS(W)	-	Word Recognition Proposed System using Word Classification

LIST OF SYMBOLS

s	-	Speech sample
$\hat{s}$	-	Predicted speech sample
a	-	Predicted coefficient
e	-	Prediction error
E	-	Mean squared error (LPC)
E	-	Energy power (Endpoint detection)
Z	-	Zero-crossing
T	-	Threshold (Endpoint detection)
D	-	Weighted Euclidean distance
R	-	Autocorrelation function
w	-	Hamming window
p	-	The order of the LPC analysis
k	-	PARCOR coefficients
c	-	Cepstral coefficients
X	-	Input nodes
Y	-	Output nodes
H	-	Hidden nodes
M	-	Weights
B	-	Bias
σ	-	Width of lattice (SOM)
λ	-	Time constant (SOM)
α	-	Learning rate (SOM)
Θ	-	The amount of influence a node's distance from the BMU (SOM)
η	-	Learning rate (MLP)
δ	-	Error information term

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

By 1990, many researchers had demonstrated the value of neural networks for important task like phoneme recognition and spoken digit recognition. However, it is still unclear whether connectionist techniques would scale up to large speech recognition tasks. There is a large variety in the speech recognition technology and it is important to understand the differences between the technologies. Speech recognition system can be classified according to the type of speech, size of the vocabulary, the basic units and the speaker independence. The position of a speech recognition system in these dimensions determines which algorithm can or has to be used. Speech recognition has been another proving ground for neural networks. Some researchers achieved good results in such basic tasks as voiced/unvoiced discrimination (Watrous, 1988), phoneme recognition (Waibel *et al.*, 1989), and spoken digit recognition (Peeling and Moore, 1987). However, research in finding a good neural network model for robust speech recognition still has a wide potential to be developed.

Why does the speech recognition problem attract researchers? If an efficient speech recognizer is produced, a very natural human-machine interface would be obtained. By natural means something that is intuitive and easy to be used by a person, a method that does not require special tools or machines but only the natural capabilities that every human possesses. Such a system could be used by any person who is able to speak and will allow an even broader use of machines, specifically computers.

1.2 Background of Study

Neural network classifier has been compared with other pattern recognition classifiers and is explored as an alternative to other speech recognition techniques. Lippman (1989) has proposed a static model which is employed as an input pattern of Multilayer Perceptron (MLP) network. The conventional neural network (Pont *et al.*, 1996; Ahkuputra *et al.*, 1998; Choubassi *et al*., 2003) defines a network as consisting of a few basic layers (input, hidden and output) in a Multilayer Perceptron type of topology. Then a training algorithm such as backpropagation is applied to develop the interconnection weights. This conventional model or system has also been used in a variety of pattern recognition and control applications that are not effectively handled by other AI paradigms.

However, there are some difficulties in using MLP alone. The most major difficulty 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 (MLP) usually needs input pattern of fixed length (Lippman, 1989). This is the reason why the MLP has difficulties when dealing with temporal information (essential speech information or feature extracted during speech processing). 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.

Since the early eighties, researchers have been using neural networks in the speech recognition problem. One of the first attempts was Kohonen's electronic typewriter (Kohonen, 1992). It uses the clustering and classification characteristics of the Self-Organizing Map (SOM) to obtain an ordered feature map from a sequence of feature vectors which is shown in Figure 1.1. The training was divided into two stages, where the first stage 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, as example, 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. Figure 1.2 shows the sequence of the responses obtained from the trained feature map when the Finnish word *humppila* was uttered. 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 (Grammatical Transformation Rules) in order to obtain words from it.

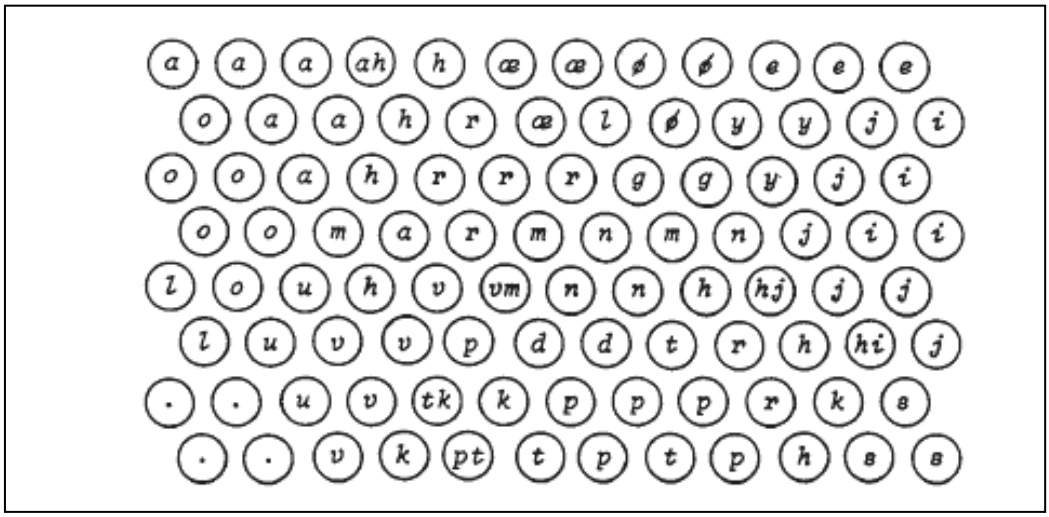

Figure 1.1: Feature map with neurons (circles) which is labeled with the symbols of the phonemes to which they "learned" to give the best responses.

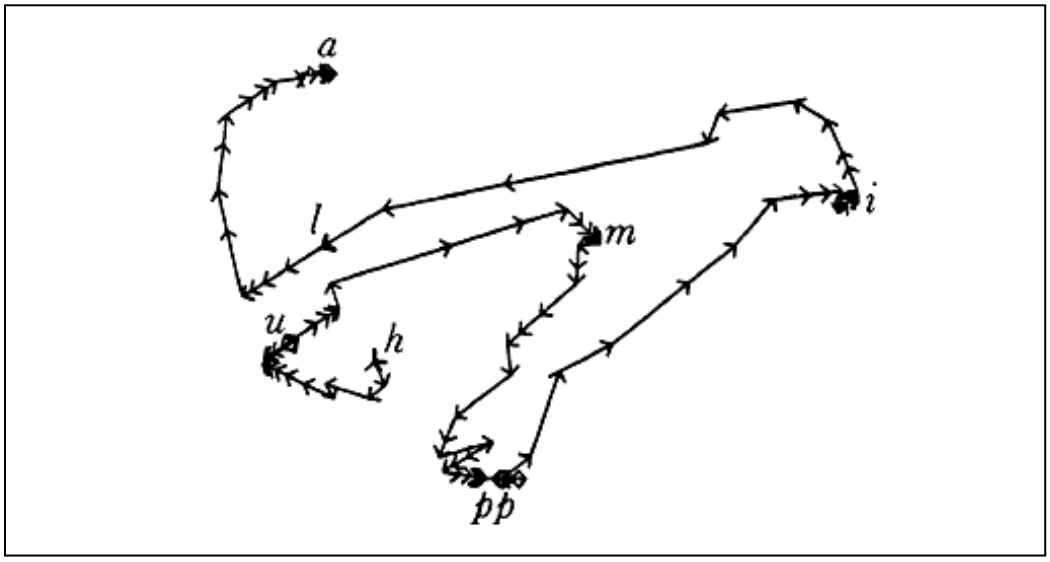

Figure 1.2: The sequence of the responses obtained from the trained feature map when the Finnish word *humppila* was uttered.

Usage of an unsupervised learning neural network as well as SOM seems to be wise. The SOM constructs a topology preserving mapping from the high-dimensional space onto map units (neurons) in such a way that relative distances between data points are preserved. The way SOM performs dimensionality reduction is by producing a map of usually 2 dimensions which plot the similarities of the data by grouping similar data items together. Because of its characteristic which is able to form an ordered feature map, the SOM is found to be suitable for dimensionality reduction of speech feature. Forming a binary matrix to feed to the MLP makes the training and classification simpler and better. Such a 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 where the utterances for same word by same speaker may be different in duration and speech rate).

1.3 Problem Statements

According to the background of study, here are the problem statements:

i. Various approaches have been introduced for Malay speech recognition in order to produce an accurate and robust system for Malay speech recognition. However, there are only a few approaches which have achieved excellent performance for Malay speech recognition (Ting *et al*., 2001a, 2001b and 2001c; Md Sah Haji Salam *et al.*, 2001). Thus, research in speech recognition for Malay language still has a wide potential to be developed.

ii. Multilayer Perceptron (MLP) has difficulties when dealing with temporal information. Since the word has to be recognized as a whole, the word patterns have to be warped using some pre-defined paths in order to obtain fixed length word patterns (Tebelskis, 1995; Gavat *et al*., 1998). Thus, an efficient model is needed to improve this drawback.

iii. Self-Organizing Map (SOM) is considered as a suitable and effective approach for both clustering and dimensionality reduction. However, is SOM an efficient neural network to be applied in MLP-based speech recognition in order to reduce the dimensionality of feature vector?

1.4 Aim of the Research

The aim of the research is to investigate how hybrid neural network can be applied or utilized in speech recognition area and propose a hybrid model by combining Self-Organizing Map (SOM) and Multilayer Perceptron (MLP) for Malay speech recognition in order to achieve a better performance compared to conventional model (single network).

1.5 Objectives of the Research

i. Studying the effectiveness of various types of neural network models in terms of speech recognition.
ii. Developing a hybrid model/approach by combining SOM and MLP in speech recognition for Malay language.
iii. Developing a prototype of Malay speech recognition which contains three main components namely speech processing, SOM and MLP.
iv. Conducting experiments to determine the optimal values for the parameters (cepstral order, dimension of SOM, hidden node number, learning rate) of the system in order to obtain the optimal performance.
v. Comparing the performance between conventional model (single network) and proposed model (SOM and MLP) based to the recognition accuracy to prove the improvement achieved by the proposed model. The recognition accuracy is based on the calculation of percentage below:

$$\text{Recognition Accuracy (\%)} = \frac{\text{Total of Correct Recognized Word}}{\text{Total of Sample Word}}$$

1.6 Scopes of the Research

The scope of the research clearly defines the specific field of the study. The discussion of the study and research is confined to the scope.

i. There are two datasets created where one is used for digit recognition and another one is used for word recognition. The former consists of 10 Malay digits and the latter consists of 30 selected two-syllable Malay words. Speech samples are collected in a noise-free environment using unidirectional microphone.

ii. Human speakers comprise of 3 males and 3 females. The system supports speaker-independent capability. The age of the speakers ranges between 18 – 25 years old.

iii. Linear Predictive Coding (LPC) is used as the feature extraction method. The method is to extract the speech feature from the speech data. The LPC coefficients are determined using autocorrelation method. The extracted LPC coefficients are then converted to cepstral coefficients.

iv. Self-Organizing Map (SOM) and Multilayer Perceptron (MLP) is applied in the proposed system. SOM acts as a feature extractor which converts the higher-dimensional feature vector into lower-dimensional binary vector. Then MLP takes the binary vectors as its input for training and classification.

1.7 Justification

Many researchers have worked in automatic speech recognition for almost few decades. In the eighties, speech recognition research was characterized by a shift in technology from template-based approaches (Hewett, 1989; Aradilla *et al.*, 2005) to statistical-based approaches (Gold, 1988; Huang, 1992; Siva, 2000; Zbancioc and Costin, 2003) and connectionist approaches (Watrous, 1988; Hochberg *et al*., 1994). Instead of Hidden Markov Model (HMM), the use of neural networks has become another idea in speech recognition problems. Anderson (1999) has made a comparison between statistical-based and template-based approaches. Today's research focuses on a broader definition of speech recognition. It is not only concerned with recognizing the word content but also prosody (Shih *et al.*, 2001) and personal signature.

Despite all of the advances in the speech recognition area, the problem is far from being completely solved. A number of commercial products are currently sold in the commercial market. Products that recognize the speech of a person within the scope of a credit card phone system, command recognizers that permit voice control of different types of machines, "electronic typewriters" that can recognize continuous speech and manage several tens of thousands word vocabularies, and so on. However, although these applications may seem impressive, they are still computationally intensive, and in order to make their usage widespread more efficient algorithms must be developed. Summing up, there is still room for a lot of improvement and research.

Currently there are many speech recognition applications released, whether as a commercial or free software. The technology behind speech output has changed over times and the performance of speech recognition system is also increasing. Early system used discrete speech; *Dragon Dictate* is the only discrete speech system still available commercially today. On the other hand, the main continuous speech systems currently available for PC are *Dragon Naturally Speaking* and *IBM ViaVoice*. Table 1.1 shows the comparison of different speech recognition systems with the prototype to be built in this research. This comparison is important as it gives an insight of the current trend of speech recognition technology.

Table 1.1: Comparison of different speech recognition systems

Software / Feature	Dragon Dictate	IBM Voice	Naturally Speaking 7	Microsoft Office XP SR	Prototype To Be Built
Discrete Speech Recognition	√	X	X	X	√
Continuous Speech Recognition	X	√	√	√	X
Speaker Dependent	√	√	√	√	√
Speaker Independent	X	X	X	X	√
Speech-to-Text	√	√	√	√	√
Active Vocabulary Size (Words)	30,000 – 60,000	22,000 – 64,000	300,000	Finite	30 – 100
Language	English	English	English	English	Malay

In this research, the speech recognition problem is transformed into simplified binary matrix recognition problem. The binary matrices are generated and simplified while preserving most of the useful information by means of a SOM. Then, word recognition turns into a problem of binary matrix recognition in a smaller dimensional feature space and this performs dimensionality reduction. Besides, the comparison between the single-network recognizer and hybrid-network recognizer conducted here sheds new light on future directions of research in the field. It is important to understand that it is not the purpose of this work to develop a full-scale speech recognizer but only to test proposed hybrid model and explore its usefulness in providing more efficient solutions in speech recognition.

1.8 Thesis Outline

The first few chapters of this thesis provide some essential background and a summary of related work in speech recognition and neural networks.

Chapter 2 reviews the field of speech recognition, neural network and also the intersection of these two fields, summarizing both past and present approaches to speech recognition using neural networks.

Chapter 3 introduces the speech dataset design.

Chapter 4 presents the algorithms of the proposed system: speech feature extraction (Speech processing and SOM) and classification (MLP).

Chapter 5 presents the implementation of the proposed system: Speech processing, Self-Organizing Map and Multilayer Perceptron. The essential parts of the source code are shown and explained in detail.

Chapter 6 presents the experimental tests on both of the systems: conventional system and the proposed system. The tests are conducted using digit dataset for digit recognition and word dataset for word recognition. For word recognition, two classification approaches are applied such as syllable and word

classification. The tests are conducted on speaker-independent system with different values of the parameters in order to obtain optimal performance according to the recognition accuracy. Discussion and comparison of the experimental results are also included in this chapter.

Chapter 7 presents the conclusions and future works of the thesis.