SPEAKER INDEPENDENT SPEECH RECOGNITION USING NEURAL NETWORK

By

TAN CHIN LUH

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia,
in Fulfilment of Requirements for the Degree of Master of Science

December 2004

Dedicated to

My beloved family and Chak Kin for their support and patience

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Master of Science.

SPEAKER INDEPENDENT SPEECH RECOGNITION USING NEURAL NETWORK

By

TAN CHIN LUH

December 2004

Chairman : Associate Professor Adznan Jantan, Ph. D

Faculty : Engineering

In spite of the advances accomplished throughout the last few decades, automatic speech recognition (ASR) is still a challenging and difficult task when the systems are applied in the real world. Different requirements for various applications drive the researchers to explore for more effective ways in the particular application. Attempts to apply artificial neural networks (ANN) as a classification tool are proposed to increase the reliability of the system. This project studies the approach of using neural network for speaker independent isolated word recognition on small vocabularies and proposes a method to have a simple MLP as speech recognizer. Our approach is able to overcome the current limitations of MLP in the selection of input buffers' size by proposing a method on frames selection. Linear predictive coding (LPC) has been applied to represent speech signal in frames in early stage. Features from the selected frames are used to train the multilayer perceptrons (MLP) feed-forward back-propagation (FFBP) neural network during the training stage. Same routine has been applied to the speech signal during the recognition stage and the unknown test pattern will be classified to one of the nearest pattern. In short, the selected frames represent the local features of the speech signal and all of them contribute to the global similarity for the whole speech signal. The analysis, design and the PC based voice dialling system is developed using MATLAB®.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Master Sains.

PENGGUNAAN RANGKAIAN NEURAL DALAM PENGECAMAN PERTUTURAN UNTUK SISTEM PETUTUR TAK BERSANDAR

By

TAN CHIN LUH

Disember 2004

Pengerusi: Profesor Madya Adznan Jantan, Ph. D

Fakulti : Kejuruteraan

Walaupun dengan kejayaan dalam pencapaian semenjak beberapa dekad yang lepas, pengecaman petuturan automatik masih merupakan satu tugas yang payah apabila ia dikait-guna dalam kehidupan harian. Keperluan yang berlainan berdasarkan aplikasi mendesak para penyelidik meninjau pelbagai cara baru demi keberkesanan sistem dalam aplikasi tertentu. Percubaan untuk penggunaan rangkaian neural tiruan (artificial neural network) sebagai alat pengkelasan dicadangkan untuk meningkat keberkesanan sistem. Projek ini menerangkan penggunaan rangkaian neural tiruan (artificial neural network) dalam sistem pengecaman pertuturan untuk perkataan berasingan. Sistem ini merupakan sistem pertutur tak bersandar (speaker independent), yang bermaksud sistem ini dapat mengecam pertuturan daripada pelbagai pengucap. Methology yang dicadangkan dapat mengatasi masalah pemilihan rangka dengan mencadangkan satu cara pemilihan rangka yang berkesan. Kod ramalan linear (linear predictive coding - LPC) digunakan untuk mewakili isyarat pertuturan dalam rangka pada takat awal. Ciri-ciri dari rangka yang dipilih digunakan untuk melatih rangkaian perceptron berlapis suap-hadapan penyebaranbalik (multilayer perceptrons feed-forward back-propagation neural network - MLP-FFBP) dalam takat pelatihan. Rutin yang sama digunakan terhadap isyarat pertuturan semasa takat pengecaman dan pola yang tidak dikenali akan diklasifikasikan ke dalam pola yang terdekat. Keseluruhannya, rangka yang terpilih mewakili ciri-ciri setempat dalam isyarat pertuturan manakala gabungan beberapa rangka mewakili ciri-ciri keseluruhan untuk isyarat petuturan. Analisa, rekabentuk dan juga pembinaan sistem dijalankan dengan menggunakan MATLAB$^{\circledR}$.

ACKNOWLEDGEMENTS

I wish to thank Dr. Adznan bin Jantan for his guidance, encouragement, and the time he spent for me during these two years over all inconvenient situations. I can’t imagine the path that I have to go through without his guidance. Besides, I also wish to thank Pn. Roslizah and Dr. Abdul Rahman Ramli for their valuable suggestions and advices towards my completion of this thesis.

Many friends around me have influenced this thesis, Soo, Khalid, and others members under ABJ Research Group, thanks all for their support and encouragement. Without them, I won’t be able to get out from the bottle neck easily. I also wish to thank Tze Kwan, for her helps in reading through my thesis.

Finally, I wish to thank my parents, for having raised me in loving environment, which has enabled me to overcome my tough time. Chak Kin, for her patience on my temper.

I certify that an Examination Committee met on 10th December 2004 to conduct the final examination of Tan Chin Luh on his Master of Science, thesis entitled "Speaker Independent Speech Recognition Using Neural Network" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Sudhanshu S. Jamuar, Ph.D.
Professor
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Mohamad Khazani Abdullah , Ph.D.
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Member)

Norman Mariun, Ph.D.
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Member)

Ali Yeon Md. Shakaff, Ph.D.
Professor
Faculty of Engineering
Kolej Universiti Kejuruteraan Utara Malaysia
(Independent Examiner)

GULAM RUSUL RAHMAT ALI, Ph.D.
Professor/Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date :

This thesis submitted to the Senate of Universiti Putra Malaysia has been accepted as fulfilment of the requirements for the degree of Master of Science. The members of the Supervisory Committee are as follows:

Adznan bin Jantan, Ph.D.
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Abdul Rahman bin Ramli, Ph.D.
Institude of Advance Technology
Universiti Putra Malaysia
(Member)

Roslizah bin Ali
Faculty of Engineering
Universiti Putra Malaysia
(Member)

____
AINI IDERIS, Ph.D.
Professor/Dean
School of Graduate Studies
Universiti Putra Malaysia

Date :

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

TAN CHIN LUH

Date :

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	artificial neural network
DFT	-	discrete Fourier transform
FFT	-	fast Fourier transform
FNN	-	fuzzy neural network
GA	-	genetic algorithm
HMM	-	Hidden Markov Model
LPC	-	linear predictive coding
LVQ	-	learning vector quantization
MFCC	-	mel-frequency cepstral coefficients
MLP	-	Multi-layer Perceptrons
RBF	-	radial basis function
RNN	-	Recurrent Neural Network
SD	-	speaker-dependent
SOFM	-	Self-organizing feature maps
SOM	-	Self Organizing Map
TDNN	-	Time Delay Neural Network
TDC	-	two-dimensional cepstrum
VQ	-	vector quantization

CHAPTER 1

INTRODUCTION

Speech and hearing have evolved as a main tool of communication among human beings. The basic building block of the speech of any language is a set of sounds named phonemes. Since early childhood, we learn the skill of this communication form naturally that we do not realize how complex the phenomenon of speech is. Even with differences in term of accent, articulation, nasality, roughness, volume, pitch, pronunciation, and speed, we are still able to interpret the speech most of the time as long as the spoken language is the language that we are familiar with.

A human brain learns a spoken language or speech unconsciously. Children learn the basic phonemes during their first year of existence. In fact, even before children understand the meaning of the speech, they are already identifying and reacting to the sounds spoken to them by their parents. Gradually, they start to learn the meaning of words and subsequently followed by the development of their vocal tract until they start to understand words and able to pronounce them correctly. Further development continues until the child is able to utter sequences of words to form complete or semi-complete sentences. It is understood that the learning of correct grammar, adaptation to different speakers and environment and even learning of different languages will continuously occur in the life span of a human being.

Due to the familiarity to spoken language, we would also hope to interact with machines via speech. Scientists and researchers are finding their ways to produce an efficient speech recognizer so that a natural human-machine interface could be invented that replace the primitive interfaces, such as keyboard and mouse for the computer. With the existence of this human-machine interface, valuable applications would come into our life to make jobs done easier and effectively. For examples, language translation machine, smart-home controller and telephone directory assistance improve the quality of human's life.

Because of the glamour of designing an intelligent machine that can recognize the spoken language, studies have been done in various fields to achieve this goal. From the process of speech production and perception in human beings to the way a human brain learns to speak and to listen, expertise and knowledge from a wide range of disciplines are required for a successful speech recognition system. Some of the disciplines that have been widely applied to solve the speech recognition problems are: signal processing, acoustics, communication theory, computer science, and pattern recognition.

Since the human brain is efficient in speech recognition, researches have been motivated to build brain-like computational methods. This fascinating research area is known as artificial neural network. The ability of storing information or knowledge in its interneuron weights makes it becomes the area of interest especially in fields related to cognitive skills. This thesis studies the application of neural network for word level speech recognition.

1.1 Speech Signal Processing and Speech Recognition

In the field of speech recognition, a large number of algorithms and methods have been proposed for different purposes. The requirement of different applications drives the researchers to develop new algorithms or improve existing methods to serve the need in different situations. For example, speaker-dependent (SD) systems which accept the speech from specific speakers are usually applied in security system. On the other hand, speaker independent (SI) recognizers are designed to recognize speech from different speakers such as speech to text engine in word processing program to replace keyboard.

To serve various applications in this field, more and more approaches have been proposed from time to time. One of the famous algorithms, the Hidden Markov Models, has been proven to be a successful statistical modelling method, especially for continuous speech recognition [1]. However, the model does suffer from some limitations that limits applicability of the technology in the real world. Attempts were made to overcome these limitations with the adoption of some new training techniques for HMM such as improved maximum model distance (IMMD) approach [2] and outlier-emphasis for non-stationary state training algorithms [3].

Artificial Intelligent approach becomes the field of interest after seeing the success of this approach in solving problems especially the classification problems [4]. The applications of artificial neural network are proposed to meet the needs of an accurate speech recognizer. For example, neural network approach to phoneme recognition [5, 6] is proposed in Japanese vowel recognition. Besides, the combination of neural networks and linear dynamic models is proven in achieving high level of accuracy in automatic speech recognition systems [7]. Another problem in speech recognition is the increase of error in presence of noise such as in a typical office environment. Some researchers propose the use of visual information such as the lip movement [8, 9]. In this case, image processing techniques and neural network are applied to capture and analyze the lip movement.

Broadly speaking, speech recognition system is usually built upon three common approaches, namely, acoustic-phonetic approach, pattern recognition approach and artificial intelligence approach .

1.1.1 Acoustic-Phonetic Approach

The acoustic-phonetic approach attempts to decide the speech signal in sequential manner based on the knowledge on the acoustic features and the relations between the acoustic features with phonetic symbols.

This approach involves two steps as mentioned above. The first step is signal segmentation and labelling. In this process, the speech signal will be separated into different segments based on the properties of the acoustic properties. For example, there are 48 sounds in English, which include 18 vowels and their combination, 4 vowel-like consonants, 21 standard consonants, 4 syllabic sounds and a phoneme referred to as a glottal stop.

This process starts with the signal analysis to analyze the spectral of the signal. This is done by using some methods such as using filter bank methods and the class of linear predictive coding (LPC) methods. Then the features of the speech signal such as nasality, frication, formant locations, voiced-unvoiced classification and ratio of high-low frequency energy are extracted form the signal. Based on the match of the features and the phonetic units, the signal is then segmented and labelled for the following processes. The phonetic is then combined to form words or sentences.

1.1.2 Pattern Recognition Approach

The pattern recognition approach, on the other hand, classifies the speech patterns without explicit feature determination and segmentation such as in the former approach.

This method starts by measuring the feature of the speech signal by using techniques as mentioned in the acoustic-phonetic approach. Besides, another good measurement algorithm is discrete Fourier transform (DFT). The features are then used as the test patterns for the training purpose of the machine, using some sample data of relevant vocabulary in the machine. The unknown signal is then passed through the pattern classification process where it will be determined to belong to which group based on the match of the unknown data with the training data set. The decision is then made base on the best match. This system is getting more and more popular and different approaches for the classification have been introduced for a better performance.

1.1.3 Artificial Intelligence approach

The artificial intelligence approach forms the hybrid system of both acoustic-phonetic approach and pattern-recognition approach.

The concept of artificial intelligence comes into place when the scientists notice that the possibility of human thinking simulation will bring the behaviour of a machine closer to the ability of a human brain. Neural network is the most popular field in artificial intelligence which has been used in classification portions of the systems. The famous networks which have been applied in this area are:

a. Multi-layer Perceptrons (MLP)

b. Self Organizing Map (SOM)

c. Time Delay Neural Network (TDNN)

d. Recurrent Neural Network (RNN)

The basic concept of neural network is to use series of simple building blocks with a simple mathematical function to form a complex combination network which is able to compute the complex nonlinear functions. The basic building block is known as “node”, analogous to the “neuron” of the biological neural model. It is the basic mathematic function such as linear function, sigmoid function or tangent function. By arranging these building blocks together, the network is able to do the parallel computation, with some local memories known as weight. This is the approach tested in the thesis.