DYNAMIC TIME WARPING FIXED-FRAME COEFFICIENT WITH PITCH FEATURE FOR SPEECH RECOGNITION SYSTEM WITH NEURAL NETWORK

RUBITA BINTI SUDIRMAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Electrical Engineering)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

AUGUST 2007

Dedication

This thesis is dedicated to

my beloved husband whom makes doa, motivates, encourage and supports me endlessly

Muhammad Noorul Anam

&

my children whom are so supportive, understanding and always inspire me

Nur A’ishah, Nur A’qila, Nurul Adilla & Luqman Yusuf

ACKNOWLEDGEMENTS

I am so proud and grateful to have a supervisor like Professor Dr. Ir. Sheikh Hussain bin Shaikh Salleh. Without any doubt, my endless gratitude goes to him for his remarkable guidance and support, I would never forget his truly sincere taught at all times. There have been so much knowledge that he continuously poured and shared without prejudice. He is my most motivator who always encouraged and backed me up when I am in a thin line, he is my absolute mentor.

My priceless appreciation and thankfulness also goes to my second supervisor, Professor Dr. Shaharuddin bin Salleh, from Mathematics Department of the Faculty of Science. He had taught and guided me in the performance optimization of the Neural Network back-propagation algorithm and in preparing the analysis part of my thesis. He is a kind of person who always keeps me working hard to achieve my goal. A dedicated teacher leads to a successful student. Thank you again.

An appreciation and gratefulness also goes to Mr. Zuraimi Yahya who is very generous in sharing his knowledge, to Dr Robiah Adnan for her statistical help during the preparation of the thesis, and to my colleagues at the Center for Biomedical Engineering who are so generous sharing their knowledge to form this success.

Thank you also to every individual in the Faculty of Electrical Engineering who are involved and had helped me in the making of this success, from the clerks to the professors, either directly or indirectly. Save the best for last, my unforgettable appreciation also goes to MOSTI and the Universiti Teknologi Malaysia who provided the financial support during my entire study. Thank you all.

ABSTRACT

Automatic Speech Recognition products are already available in the market since many years ago. Intensive research and development still continue for further improvement of speech technology. Among typical methods that have been applied to speech technology are Hidden Markov Model (HMM), Dynamic Time Warping (DTW), and Neural Network (NN). However previous research relied heavily on the HMM without paying much attention to Neural Network (NN). In this research, NN with back-propagation algorithm is used to perform the recognition, with inputs derived from Linear Predictive Coefficient (LPC) and pitch feature. It is known that back-propagation NN is capable of handling large learning problems and is a very promising method due to its ability to train data and classify them. NN has not been fully employed as a successful speech recognition engine since it requires a normalized input length. The nonlinear time normalization based on DTW is identified as the suitable tool to overcome time variation problem by expanding or compressing the speech to a desired number of data. The proposed DTW frame fixing (DTW-FF) algorithm is an extended DTW algorithm to reduce the number of inputs into the NN. This method had reduced the amount of computation and network complexity by reducing the number of inputs by 90%. Therefore a faster recognition is achieved. Recognition using DTW showed the same results when LPC or DTW-FF feature were used. This indicates no loss of information occurred during data manipulation. Pitch estimate is another feature introduced to the NN that has helped to increase recognition accuracy. An average of 10.32% improvement is recorded when pitch is added to DTW-FF feature as input to back-propagation NN using Malay digits samples. The back-propagation algorithm was then designed with both the Quasi Newton and Conjugate Gradient methods. This is to compare which method is able to achieve optimal global minimum. Results showed that Conjugate Gradient performed better.

ABSTRAK

Produk sistem pengecaman pertuturan otomatik sudah banyak terdapat di pasaran sejak beberapa tahun yang lalu. Penyelidikan dan pembangunan intensif masih dilakukan untuk menambahbaik teknologi pertuturan. Antara kaedah pengecaman yang digunakan dalam teknologi pertuturan adalah Model Markov Tersembunyi (HMM), Bengkukan Masa Dinamik (DTW) dan Rangkaian Neural (NN). Namun kaedah HMM lebih mendapat perhatian dan sebaliknya bagi teknik rangkaian neural. Dalam penyelidikan ini kaedah rangkaian neural dengan algorithma perambatan-balik digunakan untuk pengecaman, di mana masukan terdiri daripada ciri pekali ramalan lelurus (LPC) dan pic. Algorithma perambatan-balik telah diketahui beupaya mengendalikan masalah pembelajaran yang besar dan rumit, namun ia berkebolehan melatih dan mengkelaskan data. NN belum berjaya sepenuhnya apabila digunakan sebagai enjin pengecaman kerana ia memerlukan masukan data yang sama kepanjangan. Pernormalan masa tidak linar berlandaskan DTW dikenalpasti sebagai kaedah yang sesuai digunakan untuk mengatasi masaalah perbezaan tempoh. DTW berupaya memanjang atau memendekkan tempoh sesuatu sebutan perkataan kepada bilangan bingkai masa data yang dikehendaki. Kaedah Penyesuaian Bingkai DTW (DTW-FF) yang telah dicadangkan berupaya mengurangkan bilangan masukan ke rangkaian neural. Kaedah ini dapat mengurangkan pengiraan dan kekompleksan rangkaian dengan pengurangan bilangan masukan sebanyak 90%. Oleh itu, pengecaman adalah lebih pantas. Pengecaman menggunakan DTW dengan masukan LPC dan pekali DTW-FF masing-masing memberikan keputusan yang sama. Ini menunjukkan tiada informasi hilang ketika manipulasi data. Anggaran pic adalah satu lagi ciri diperkenalkan ke dalam NN yang telah membantu menaikkan tahap ketepatan pengecaman. Purata 10.32% kenaikan direkodkan apabila pic digandingkan dengan ciri DTW-FF sebagai masukan ke rangkaian neural berdasarkan sampel digit Melayu. Algorithma perambatan-balik direkabentuk menggunakan kedua-dua kaedah *Quasi Newton* dan *Conjugate Gradient*. Ia bertujuan untuk membandingkan teknik mana yang berupaya mencapai titik global optima yang minimum. Keputusan menunjukkan kaedah *Conjugate Gradient* adalah lebih baik.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Description
F_o^r	raw pitch frequency
F_o^o	optimized pitch frequency
α	momentum rate
ε	error
ℓ	length
τ	cycle period
δ	delta
β	conjugate gradient constant
λ	wavelength
$\eta\varepsilon$	learning rate
A_k	tube area function
c	speed of light, 35000 cm/sec
d	Local distance
D	Global distance
F^-	frame compression
F^+	frame expansion
$F0$	fundamental frequency

f_s	sampling frequency
M, N	window length
r	reflection coefficient
Re	Reynold’s number
U_G	glottis volume velocity
U_L	lips volume velocity

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Speech recognition describes a group of special technologies that allow callers to speak words, phrases, or utterances that are used to control some particular applications. In the case of voice processing, speech recognition is used to replace touch-tone input, make for more intuitive menu structures, and add a level of simplicity and security to some systems. Speech recognition, on the other hand, is a technology that uses the spoken word as input that has an effect on the logic flow and execution of the program in query.

Speech recognition has witnessed so many approaches in dealing with the matter. Many Automatic Speech Recognition (ASR) systems have been developed. Among them are DRAGON, SONIC and SPHINX. Most of the systems are based on the state-of-the-art Hidden Markov Model (HMM) method or combination of HMM and Artificial Neural Network (ANN). HMM is a dominant technology used in ASR in which it works based on likelihood estimation of each phoneme. Early studies have used small vocabulary isolated words and since then the studies have been extended to continuous speech and large vocabulary system. Many systems have been using hybrid methods like HMM/ANN, ANN/ HMM, Dynamic Time Warping (DTW) and Multi-Layer Perceptron (MLP), and some are using Time Delay Neural Network (TDNN) for the same purpose.

In this chapter, the problem background is introduced. It is followed by the objective and stretches to the scopes of study. The literature review continues after that and then followed by the brief research methodologies, contribution of the thesis. The last section in this chapter is the organization of the thesis.

1.2 Problem Statement

Today, speech recognition can be considered as a mature technology, where current research and technologies have complex combinations of methods and techniques to work well with each other towards the refinement of the recognition. If for instance a neural network wanted to be used as the recognizer, one would intend to have a method that can reduce the network complexity with less storage requirement which in return it will give faster recognition, therefore that method has to be formulated. In that respect, a formula using combination of methods available need to be constructed, or create a new form of feature that can represent the speech information without losing much of their important information. The intended method must provide a recognition performance at least not less than the best existing method. For this purpose, time normalization (also known as time alignment) method using a non-linear time alignment namely DTW is investigated and feature vectors manipulations are performed to suit the back-end proposed recognition engine, which is back-propagation neural network algorithm.

The issue that wanted to be addressed in this study is the time normalization problem, in which it is a critical problem that has to be tackled if speech recognition wanted to be performed. DTW was selected and used as the time normalization technique to obtain uniform speech lengths due to its least complexity and fast computation. A new method, which is based on the DTW called as DTW Frame Fixing (DTW-FF) algorithm was proposed to extract another form of feature but yet it is still being able to keep the original information contains in the speech signals.

As pitch is a good speech feature when coupled with other feature, it will be used in combination with the new feature extracted from the DTW-FF algorithm into the neural network speech pattern recognition. Their recognition performance is investigated and compared to previous works done which used pitch and other feature combination.

An optimization to the network is also performed to avoid the global minimum from being trapped in a local minimum valley when using the back-propagation algorithm with the steepest gradient descent search method. In this study, the network is designed to handle the parallel processing of multiple samples/words. Therefore it caused the network to compute a large amount of connection weights as well as the error updates at a time. As a consequence a longer time is taken for the network to converge to its global minima. Since the Conjugate Gradient method has been proven of being able to accelerate the network convergence, it is applied into the back-propagation mechanism to replace the steepest gradient descent algorithm. The Quasi-Newton method is also tested as a comparison with regard to their convergence performances.

1.3 Objective of Study

The main objective of this research is to find an alternative method to reduce the amount of computation and complexity in a neural network. In this case it is for speech recognition so that the rate of convergence can be increased. One of the ways to achieve this objective is by reducing the number of inputs into the network. This is done through dynamic warping process in which local distance scores of the warping path are utilized as the network's input instead of the global distance scores.

In this study, the extracted pitch feature firstly is optimized using pitch-scaled harmonic filter (PSHF) algorithm to reduce glitches during the voice activity before it is used as input into the Neural Network (NN).

Then, at the last stage, optimization is performed to the back-propagation algorithm of the neural network using Quasi-Newton and Conjugate Gradient methods to search for the fastest convergence rate between these two techniques.

The current method of back- propagation is based on the steepest gradient descent method as its searching direction, in which this method is exposed to a bad local minima settlement. Another aspect leading to the optimization was that a large number of inputs is presented into the NN.

1.4 Scope of Study

In order to achieve all of the objectives set above, several scopes have been outlined. The scopes of the study are:

i. The study utilized only samples of Malay digits 0-9 which were uttered by 11 speakers (6 males and 5 females), uttered 5 times in each session for 5 different sessions.

ii. The feature extraction technique selected was the linear predictive method.

iii. The time normalization method was based on dynamic time warping principles.

iv. The pattern recognition engine was using the back-propagation neural network. Preliminary experiments were conducted to find the NN parameters used in the study using the speech samples collected to fit the network requirements.

v. The pitch feature was also investigated along with the DTW-FF coefficients to determine if using pitch could improved the recognition performance, especially to the subjects that showed low recognition percentage.

vi. An optimization to the back-propagation error search was used to investigate the best method that can avoid the bad local minima (error is at its optimal global minimum).

1.5 Literature Review

Speech recognition has gained a wide attention from researchers in the field. This has been going on since the past half decade and still going on to search either for a better solution to existing equipments or methods, or to find alternative ways of dealing with particular problem that are still looking for solutions [Lippmann, 1989; Salleh, 1997; Nong *et al*., 2002; Markov and Nakamura, 2003]. It includes the pitch contribution to speech recognition or methods used for dealing with time variations in uttered speech. Many methods of speech recognition present, among the popular ones are Dynamic Time Warping (DTW), Neural Network (NN) [Liu *et al.*, 1992; Kuah *et al*., 1994; Lee *et al*., 1998] and Hidden Markov Model (HMM). However, these methods have their own strengths and weaknesses. DTW is considered as one of the popular methods due to its low cost matching technique and is good especially for isolated words recognition, so does NN [Demichelis *et al.*, 1989; Matsuura *et al.*, 1994; Tabatabaee *et al*., 1994; Seddik *et al*., 2004]. HMM requires a lot of computations and storage, but it is good either for continuous or large vocabulary speech recognition [Lee, 1990; Juang and Rabiner, 1991; Zhao, 1993; Woodland *et al*., 1994].

Speech recognition has been the center of attention for many researchers in the fields more than a century ago. Until today speech recognition is still gaining interests from researchers as it is one of the complex problems to solve either in the feature extraction or in pattern recognition. Front end processing of speech data or frequently called as feature extraction, typically utilizes linear predictive coefficient method (LPC), mel-frequency (MFCC), or spectral method such as the fundamental frequency, formants, or power spectral density. A series of vectors characterizing the time varying spectral feature of speech signal is the result of feature extraction. Methods of recognition have been mainly using Hidden Markov Model (HMM), Dynamic Time Warping (DTW), and Neural Network (NN). HMM is a method which consider the probability of sample occurrence in different states and requires initial assumption to predict the next state's probability. DTW is a method in which the recognition is done based on the time warping path between sample and reference signals.

There are a number of approaches in solving the problem of automatic speech recognition from the past until present:

- Sakoe and Chiba (1978) are the first to introduce dynamic programming into DTW. This technique is guaranteed to find the least distance path between two signals, while it also minimizes the amount of computation.
- Abdulla *et al.* (2003) has used MFCC to prepare reference template for DTW based speech recognition particularly for small vocabulary.
- Lippmann (1989) and Salleh (1997) reviewed and summarized the works done by previous researchers who were using neural networks and HMM in speech recognition. The best speech recognizers mostly perform well in an artificial constrained task. HMM is considered as the current best method. However it has the limitation where its modeling in low-level and high-level is poor. In this respect, neural network is seen as having the potential of overcoming these problems.
- Botros *et al*. (1992) used DTW and MLP with sequence of dynamic networks, they did not perform time alignment using DTW, but they only used DTW to find the global distance score and used that score as the input into their MLP.
- Jang and Un (1996) used HMM combined with NN.
- Chen et al. (1996) used DTW just to time-aligned the input pattern and use MLP as the recognizer utilizing the total distance score. They also used other combinations like MLP/HMM and HMM/MLP.
- Other works involved DTW and NN also include Prasanna *et al*. (2004), he and his colleagues used DTW as feature extractor for spectral and suprasegmental features. They used the warping path total distance scores (while mentioning that they ignored the warping path information) for both extracted features for their text-dependent speaker verification system.
- A work by Soens and Verhelst (2005) used split time warping to improve automatic time synchronization of speech. This is a kind of time alignment or normalization to two similar speeches in which their duration are varied. It is the emphasis of this study to normalize or time aligned these two varied similar speeches. However, Soens and Verhelst (2005) utilized warping

algorithm based on timing reference, in which their method made insertion and deletion in the reference and unknown source when either one of the sources showed a pause (silence), not only into the reference or unknown source itself but the insertion or deletion is also done onto both of the sources. This eventually increased the number of frames or coefficients to represent the speech. On the other hand, in this study, the frame compression and expansion are done within the reference utterance which is based on the frames count of the reference template, not both sources (the reference and the unknown input source).

The method using DTW has also been used in Abd-Aziz (2004), in which the technique is called the cross match model, in which it is based on cross correlation between two speech tokens. The aim is to find the similarity of two speech tokens by which they are correlated by their correlation factor. The features used are the LPC and MFCC for vector quantization (VQ). Ariff *et al.* (2005) used discrete HMM as classifiers for speaker recognition using isolated Malay digits database, their result also had shown a good recognition using the LPC feature with a majority of above 90%.

Neural networks can be efficient in dealing with speaker recognition because they have powerful discrimination abilities, but neural network cannot deal fully with the problem of time variability of speech [Salleh, 1997]. This problem can be handled using DTW and HMM method efficiently. In this study, this problem is tackled using the DTW method before the data can be presented to the neural networks for recognition. Neural networks back propagation is chosen as the recognition engine due to its ability to discriminate the classes of data efficiently as compared to other method like HMM. Previous studies had shown the hybrid methods of HMM and NN, DTW[1] and HMM, as well as DTW[2] and NN as reported by Lippmann (1989).

After looking at different methods or hybrid methods used in the speech recognition subsets, other avenue of speech recognition can be investigated. An

[1&2] The DTW method was utilizing the global distance score as the input coefficient.

example is the hybrid method introduced in this study, it consists of the DTW method (this time the study utilized the local distance score) and multilayer neural networks. Multilayer network is chosen because its design suitable for pattern association, i.e, mapping input vectors to output vectors. Recurrent networks are useful for pattern sequencing, i.e, following sequences of network activation over time, while modular networks are useful for building complex systems from simpler components.

Time normalization is a typical method to interpolate input signal into a fixed size of input vector. The linear time normalization is the simplest method to overcome time variation, but it is a poor method since it does not take into account some important feature vectors when deleting or duplicating them to shorten or lengthen the pattern vectors, if required [Creaney-Stockton, 1996]. Nonetheless, since then it has been the fundamental method for compression and expansion for speech pattern vector.

The non-linear methods are more complex than the linear ones but they take into account the importance of the feature vectors when they are manipulated to change the length of the speech pattern. The non-linear methods are time warping and trace segmentation; they are more suitable to carry out the intentions of fixing the input vector to a specified size. Indirectly, the pre-processing also applies trace segmentation method, in which the idea of trace segmentation is to reduce the number of stored feature vectors for the stationary portion during the speech [Cabral Jr. and Tattersall, 1995]. In other words trace segmentation method is also a subset of dynamic warping method. Unfortunately, trace segmentation is not fully used as the normalization technique because it does not provide good performance even compared to DTW. The technique shared a common compression technique, but not the expansion when compared to DTW. Furthermore, the distance segmentation is inappropriate as well as the spatial sampling rate along the trace [Cabral Jr. and Tattersall, 1995], plus it can only perform frame reduction during the stationary speech portion. The idea in this thesis is to reduce the number of stored feature vectors from combination methods of trace segmentation method and the DTW technique.

Even though research in DTW has been used since more than a decade ago, it is still being explored in different angles of interest. However, only non-linear time normalization technique can perform both frame expansion and reduction, and still can preserve the important features during the process [Salleh, 1993; 1997]. In that sense, DTW time normalization is selected and used to obtain uniform speech lengths and use them in the pattern recognition stage later. From the literature reviews, past and most current research are using the global distance scores as a measure [Chen *et al*., 1996; Tsai and Lee, 1997; Abdulla *et al*., 2003; Prasanna *et al*., 2004; Soens and Verhelst, 2005], or LPC coefficients as the input to the neural network one sample at a time. In that respect, a new method called dynamic time warping frame fixing (DTW-FF) is proposed to extract another form of feature which has a smaller number of input size so that it can reduce the amount of computation and network complexities in the back-propagation neural network. Therefore, DTW-FF is a modified method that keeps the time alignment of the speech signals.

Neural network is chosen as the back-end recognition engine due to its past good and reliable performances in speech recognition. As mentioned in the earlier paragraphs, NN is considered as one of the popular method used especially when dealing with isolated word speech recognition. Since this study considers mainly on isolated words, NN is chosen as an engine to perform the recognition task. The main task of the study which is to find an alternative way of reducing the number of inputs into the NN should introduce a new form of input representation into the network, which is simpler and smaller when compared to using the LPC feature.

Realizing the approaches that had been used in the past research, the issues to be tackled were based on the total distance score (also known as global distance score) or LPC coefficients as inputs into the neural network while possibly improving the recognition performance. In this research, NN is employed as the recognition engine utilizing DTW scores obtained from frame matching algorithm. The new algorithm is based entirely on the typical DTW algorithm that utilizes dynamic programming which was introduced by Sakoe and Chiba (1978). NN with back-propagation algorithm is chosen due to its ability to reduce the recognition percentage error even though the training time is longer. However, utilizing the local distance scores obviously can reduce the input size into the NN and this should increase the training speed, thus less time and faster recognition could be achieved, i.e, faster convergence rate.

In this particular research work, combination of DTW/NN back-propagation algorithm utilized DTW to normalize all input patterns with respect to the template pattern. The frame matching (normalization) is performed to obtain the new feature representation for the recognition using the dynamic programming. In addition to that, the local distance scores are used as inputs into the MLP neural network, instead of using the global distance score like previous works have done. Using only the global distance score certainly gives less recognition percentage compared to using the local distance scores as proposed in this thesis. It is because the local distance scores represent more detail information about the speech signal along the warping path than the global distance score. Furthermore, a global distance score of different paths could be the same for different speech signals in which sometimes it was not the correct score for a particular word, thus yield to an improper recognition. Also, in this study, a hybrid method is used when the DTW-FF algorithm is coupled with the back-propagation neural network to perform the recognition.

Suprasegmental feature or also called as the voice source characteristic, namely pitch is a perceptual quantity which is introduced into the neural network as another input feature along with the DTW-FF feature. This is because LPC feature vectors itself sometimes does not give an overall high percentage of recognition, nevertheless pitch feature itself does not give high recognition rate indeed. Studies had shown that pitch is at least coupled with another feature if it wanted to be used as a feature for speech recognition [Chan *et al*., 1994; Wong and Siu, 2002; Markov and Nakamura, 2003].

Traditionally automatic speech recognition is based on some derived features which represent the vocal tract system characteristics, and leaving the knowledge of voice source characteristics, namely, as pitch. This is because pitch is not an ideal source of information for automatic speech recognition [Magimai-Doss, 2003]. This is supported by an earlier study by Fujinaga *et al.* (2001) when they found that pitch cannot work well with the combination of MFCC using the HMM. However, their recent studies supported by Stephenson *et al.* (2002) showed otherwise. They found that additional pitch frequency information can improve the performance of automatic speech recognition system which is further examined in their latest work presented in Stephenson *et al.* (2004) using the hybrid HMM/ANN.

Other works involving pitch as an input feature include:

- Singer and Sagayama (1992) used pitch for phone modeling for HMM speech recognition; their work showed that the use of pitch information consistently improves the recognition performance. They used the 16th order cepstrum coefficients with triangular regression window to extract the pitch (called as the lag-window method) which finally take the pitch in logarithm form.
- Chan *et al.* (1994) used pitch along with the first three formant frequencies, age, percentage English is used during typical day, and number of years English has been used as inputs into the neural networks for assisting ASR system in which different English accent might be used by the native and nonnative English speakers. Their results are quite promising and recommended for improvement of an ASR system. However the pitch extraction method used is not mentioned in their paper.
- Wong and Siu (2002) used tone related feature (pitch) along with MFCC feature using HMM for Chinese speech recognition.
- In other works related to the use of pitch in speech recognition is Markov and Nakamura (2003). They used the feature along with the speaker's gender into their HMM/Dynamic Bayesian Networks (BN) for isolated word recognition.

An application that strictly requires pitch information into the system is the cochlea implant; the implant device is a custom design device which only suits a particular patient because each patient has different amounts of pitch and periodicity information (which determines the fundamental frequency, *F0* of a speech).

1.6 Research Methodology

In this research, NN is employed as the recognition engine utilizing DTW scores obtained from frame matching algorithm, which based entirely on the typical DTW algorithm. NN with back-propagation algorithm is chosen due to its ability to reduce recognition percentage error even though the training time is longer. However, utilizing the local distance scores reduce the input size to NN compared to LPC coefficients and this should result to a faster training time, thus faster recognition.

Briefly, there are two features that are taken into account in this research. The first one is the local distance scores feature and the second is the pitch feature. Those are the input features to the back-propagation error algorithm of neural network, which is the focal point of this speech recognition research.

The local distance score feature goes through few processes before they are ready for NN. Firstly, after start and end point detection, features are extracted using LPC method. Then using dynamic time warping, speech is warped to a reference sample; unknown speech that being warped has varying number of frames with respect to their reference sample. However, NN requires the same length of data as their input, so prior to the recognition process, input data has to be aligned due to variation in the speech durations. Some methods have to be applied to the data so that those data are fit for the NN and that method is called the DTW-FF method. In this method, the speech signals are compared to a selected reference which is chosen based on its average frame value over a sample population, to align the signal according to their frames based on the compression and expansion technique by means of combination of trace segmentation and dynamic time warping. The DTW path type I is used whereby either one of the three slope conditions is applied to the warping path each time the path is warped. More details on the method are presented in Chapter 3 for the feature extraction part, further in Chapter 5 for the method's implementations.

The vibration of vocal cords will cause the production of speech in which the speech contains two types of acoustic information [O'Shaughnessy, 1987]. They are the voice source information and the vocal tract system information. The voice source characteristics like pitch is a perceptual quantity, actually pitch is referred to as the rate of vocal cord vibrations which can be estimated from the speech signal itself. Pitch contains a lot of information such as information about the speaker, it can tell whether the sound is a voiced or unvoiced, as well as it contains prosodic information. In our study, the second feature, which is pitch, is extracted using a method called pitch scaled harmonic filter (PSHF). In PSHF, pitch is optimized and the pitch feature is retained and used as one of the input features into the NN. These pitch features represent the formant frequencies of the spoken utterance. Referring to the literature review section, pitch has been used as one of the feature using cepstrum extraction method (MFCC feature) and most work were carried out using the HMM or combination of hybrid model of HMM/ANN or HMM/Bayesian Networks [Chan *et al.*, 1994; Fujinaga *et al.*, 2001; Stephenson *et al.*, 2002, 2004; Markov and Nakamura, 2003]. However, in this study pitch is extracted and optimized using a scaled harmonic filter algorithm and they are used into Back-Propagation NN along with another feature called the DTW-FF feature.

The last part of the study is the network performance optimization. The optimization involves the usage of Quasi-Newton method and Conjugate Gradient algorithm in place of the steepest gradient descent algorithm in the back-propagation part of the neural network. This method is able to reduce the number of iterations for error calculations and weight updates, thus improves the convergence rate.

In summary, the implementation used the following algorithms to perform the research study:

i. LPC (Linear Predictive Coding): used for feature extraction
ii. DTW-FF (Dynamic Time Warping Fixed Frame) Algorithm: used for frame fixing/ time alignment
iii. PSHF (Pitch Scaled Harmonic Filter): used for pitch feature extraction and optimization
iv. BPNN (Back-Propagation Neural Network): used as the recognition engine
v. The Quasi-Newton and Conjugate gradient method: used for network optimization.

After the addition of pitch feature to DTW-FF feature and tested using the NN, the network optimization follows. In the optimization part, which is the last scope of the study, it will compare the use of different search direction procedure for faster convergence rate, which are between the Quasi-Newton and Conjugate Gradient Method versus the Steepest Gradient Descent Method (used in the first phase of the experiment). The experiments are divided into three phases which are described in detail in Chapter 3.

1.7 Contribution of the Thesis

The main contribution of this thesis is the development of an extended dynamic time warping algorithm for the purpose of speech feature extraction for the usage in the parallel processing based on the neural network algorithm for speech recognition. The modified algorithm is called as the dynamic time warping fixed frame algorithm (DTW-FF) whereby the algorithm is able to perform speech frame compression and expansion based on the rules set in the warping path. This algorithm is important in solving the time variation problem especially during a parallel processing where multiple classes of input are used.

The second contribution of the thesis is the new combination of features obtained from the DTW-FF algorithm and PSHF (which is a pitch extraction algorithm). The PSHF algorithm is able to optimize the pitch of the speech so that an optimal pitch value is used for the input into the neural networks speech recognition. The pitch optimization is needed due to octave errors during the windowing activities. In this study, the combination of DTW-FF feature and optimized pitch feature has given very good speech recognition results despite of the pitch itself which cannot give a good representation of speech information if it was being used alone in speech recognition.

In order to improve the existing performance of the ASR system, network optimization has been implemented. The current method of NN was using the Steepest Gradient Descent search method to search for the optimal global minimum, but the steepest gradient descent method is exposed to false global minimum. Other methods that are less exposed to false global minimum are the Quasi-Newton and Conjugate Gradient methods. Experiments were performed using these three methods to compare the convergence rate according to their epochs and their global minimum sum squared error. Results have shown that Conjugate Gradient method outperformed other search methods to converge at the most optimal global minimum.

There are six international conference papers presented and three national journal paper have been published resulting form the work carried out in the study. The publication list is attached in Appendix E.

1.8 Thesis Organization

This thesis is divided into eight chapters. Chapter 1 describes the problem background and literature review that has been done, states the objective of the research, and stretches the scopes of study. The general approach of the research is also briefly presented in this chapter.

Chapter 2 explains the speech production mechanisms, illustrated with two figures of human vocal tract: general figure and anatomy of the human vocal tract. Also more elaborate definition of fricatives, especially voiced fricatives. Reviews on modeling the vocal tract for voiced fricatives are discussed. This chapter also discusses the aerodynamic and acoustic consequences to the voiced fricative speech and this lead to the importance of pitch feature in speech.

Chapter 3 discusses feature extraction and preprocessing of the features. LPC feature extraction is described in details here as well as the dynamic time warping improved algorithm or called as DTW Frame Fixing (DTW-FF) Algorithm to obtain another feature from it, which is called as DTW-FF feature.

Chapter 4 describes the neural network theory and architectures. Details on the back-propagation algorithm are described, the network limitations, and also DTW-FF feature fittings into the neural network.

Chapter 5 describes the methodology and experiments setup conducted in this study in great details. These include the adaptation of warping process as the fixing moderator.

Chapter 6 is about the discussions and analysis of the results from the conducted experiments. Step by step experiments results are presented and discussed in such a way of tackling the problem statements.

Chapter 7 discusses the needs of performance optimization to the neural network learning algorithm. Two methods of optimization to replace the steepest gradient descent in the back-propagation part are discussed in this chapter.

Chapter 8 is the last chapter, which conclude the thesis findings and also state some recommendations to go about the research in the future with the availability of the developed software like the DTW-FF and the PSHF. The software can also be used for speech synthesis rather than focusing only at speech recognition. Other features could also be considered rather than only using the LPC coefficients during the frame fixing.

REFERENCES AND BIBLIOGRAPHIES

Abdul-Aziz, M. A. (2004). *Speaker Recognition System Based on Cross Match Technique*. Universiti Teknologi Malaysia: Master Thesis.

Abdulla, W. H., Chow, D., and Sin, G. (2003). Cross-Words Reference Template for DTW-based Speech Recognition System. *IEEE Technology Conference (TENCON)*. Bangalore, India, 1: 1-4.

Ahmadi, M, Bailey, N. J., and Hoyle, B. S. (1996). Phoneme Recognition using Speech Image (Spectrogram). *Proceedings of International Conference on Signal Processing*. 1: 675-677.

Ariff, A.K. and Salleh, S. H. (2005). Digit Recognition System Based On Discrete HMM. International Conference on Robotics, Vision and Signal Processing.

Badin, P. (1991). Fricative Consonants: Acoustic and X-Ray Measurements. *Journal of Phonetics*. 19: 397-408.

Badin, P., Shadle, C. H., Ngoc, Y. P. T., Carter, J. N. , Chiu, W. S. C. , Scully, C., and Stromberg, K. (1994). Frication and Aspiration Noise Sources: Contribution of Experimental Data to Articulatory Synthesis. *Proceeding International Conference on Spoken Language Processing (ICSLP)*. 1: 163-166.

Bendat, J. S. and Piersol, A. G. (1984). *Random Data: Analysis and Measurement Procedures*. New York: Wiley Intersciene.

Botros, N. M. and Premnath, S. (1992 June). Speech Recognition using Dynamic Neural Networks. *International Joint Conference in Neural Network*. 4: 737-742.

Cabral Jr., E.F. and Tattersall, G. D. (1995 May). Trace-Segmentation of Isolated Utterances for Speech Recognition. *International Conference on Acoustics, Speech, and Signal Processing* 1:365-368.

Chan, M.V., Feng, X., Heinen, J.A., and Niederjohn, R.J. (1994). Classification of Speech Accents with Neural Networks. *IEEE International Conference on Neural Networks*. 7: 4483-4486.

Charalambous, C. (1992). Conjugate Gradient Algorithm for Efficient Training of Artificial Neural Networks. *IEE Proceedings-G*. 139(3): 301-310.

Chen, W-Y., Chen, S-H., and Lin, C-J. (1996). A Speech Recognition Method Based on the Sequential Multi-Layer Perceptrons. *Journal of Neural Networks*. 9(4): 655-669.

Cichocki, A. and Unbehauen, R. (1993). *Neural Networks for Optimization and Signal Processing*. Stuttgart: John Wiley and Sons.

Coker, C. H. and Krane, M. H. and Reis, B. Y., and Kubli, R. A. (1996). Search for an Unexplored Effects in Speech Production. *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, 14(6): 415-422.

Creaney, M. J. and Gorgui-Naguib, R. N. (1994). A Scaly Artificial Neural Network for Speaker Independent Isolated Word Recognition using Non-Linear Time Alignment. *International Conference of IEEE World Congress on Computational Intelligence*. 7: 4431-4436.

Creaney-Stockton, M. J. (1996). *Isolated Word Recognition using Reduced Connectivity Neural Networks with Non-Linear Time Alignment Methods*. University of New Castle-Upon-Tyne: Ph.D. Thesis.

Crow, S. C. and Champagne, F. H. (1971). Orderly Structure in Jet Turbulence. *Journal of Fluid Mechanics*. 48: 547-591

Crystal, T. H. and House, A. S. (1988). Segmental Duration in Connected-Speech Signals: Current Results. *Journal of Acoustical Society of America*. 83: 1553-1573.

Demichelis, P., Fissore, L., Laface, P., Micca, G., and Piccolo, E. (1989 May). On the Use of Neural Networks for Speaker Independent Isolated Word Recognition. *Proceedings of ICASSP*. 1: 314-317.

Denes, P. B. and Pinson, E. N. (1973). *The Speech Chain: The Physics and Biology of Spoken Language*. New York: Anchor Science.

Fackrell, J. W. A. (1996). *Bispectral Analysis of Speech Signals*. Department of Electrical Engineering, University of Edinburgh, UK: Ph. D. Thesis.

Fant, C. G. M. and Pauli, S. (1974). Spatial Characteristics of Vocal Tract Resonance Modes. *Proceeding of the Speech Communication Seminar*. 74: 121-132.

Fant, C. G. M. (1960). *The Acoustics of Speech Production*. The Hague, Netherlands: Mouton.

Farrell, K. R., Mammone, R. J., and Assaleh, K. T. (1994). Speaker Recognition using Neural Networks and Conventional Classifiers. *IEEE Transactions on Speech and Audio Processing*. 2(1): 194-205.

Fenglei, H and Bingxi, W. (2000). An Integrated System for Text-Independent Speaker Recognition using Binary Neural Network Classifiers. *International Conference on WCCC-ICSI – Proceedings of Signal Processing*. Beijing, China, 2: 710-713.

Fitch, H. L., Kupin, J. J., Kessler, I. J., and DeLucia, J. (2002). Relating Articulation and Acoustics Through a Sinusoidal Description of Vocal Tract Shape. *Journal of Speech Communication*. 39: 243-268.

Flanagan, J. L. and Ishizaka, K. (1976). Automatic Generation of Voiceless Excitation in a Vocal Cord Vocal Tract Speech Synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 24(2): 163-170.

Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception*. 2nd Edition. Berlin, Germany: SpringerVerlag.

Fujinaga, K., Nakai, M., Shimodaira, H., and Sagayama, S. (2001). Multiple Regression Hidden Markov Model. *International Conference on Speech and Signal Processing*. 513-516.

Gabelman, B. and Alwan, A. (2002). Analysis by Synthesis of FM Modulation and Aspiration Noise Components in Pathological Voices. *Proceeding of IEEE ICASSP*. 449-452.

Gish, H. and Schmidt, M. (1994 October). Text-Independent Speaker Identification. *IEEE Signal Processing Magazine*. 18-32.

Glaeser, A. (1996). Compact Modular Neural Networks in a Hybrid Speaker Independent Speech Recognition System. *IEEE International Conference on Neural Networks*. 4: 1895 – 1899.

Gray, H. and Lewis, W. H. (1959). *Anatomy of the Human Body*. 20th Edition. Philadelphia: Lea & Febinger.

Hackbarth, H. and Mantel, J. (1991 July). Modular Connectionist Structure for 100-Word Recognition. *International Joint Conference on Neural Networks*. Seattle, 2: 845-849.

Hagan, M.T., Demuth, H.B., and Beale, M. (1996). *Neural Network Design*. Boston: PWS Publishing Company.

Harb, H. and Husseiny, A. H. (2000 December). Isolated Word Recognition using Neural Networks. *The 7th IEEE International Conference on Electronics, Circuits, and Systems*. 1:349-351.

Harris, K. S. (1958). *Cues for the Discrimination of American English Fricatives in Spoken Syllables*. Language and Speech Series. Robert Draper Ltd. 1:15.

Hattori, H. (1992). Text-Independent Speaker Recognition Using Neural Network. *Proceedings of ICASSP*. 2: 153-156, San Francisco.

Holmes, J. and Holmes, W. (2002). *Speech Synthesis and Recognition*. 2nd Edition. London: Taylor and Francis.

Huckvale, M. A. (2003). *Speech Filing System SFS, 2003*. Release 4. 4. Department of Phonetic and Linguistic, University College London, UK. http://www. phon. ucl. ac. uk/resource/sfs/

Jackson, P. J. B. (2001). Acoustic Cues of Voiced and Voiceless Plosives for Determining Place of Articulation, *Proceeding of Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC)*. Aalborg, Denmark. 19-22.

Jackson, P. J. B. and Mareno, D. (2003). *PSHF Beta Version 3*. 10, CVSSP – University of Surrey, Guilford, UK. http://www.ee.surrey.ac.uk/Personal/P.Jackson

Jackson, P. J. B. and Shadle, C. H. (2000). Frication Noise Modulated by Voicing as Revealed by Pitch-Scaled Decomposition. *Journal of Acoustical Society of America*. 108(4): 1421-1434.

Jackson, P. J. B. and Shadle, C. H. (2001). Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence Noise Components in Speech. *IEEE Transactions on Speech and Audio Processing***.** 9(7): 713-726.

Jiang, M., Pang, H., Deng, B., and Zong, C. (2004). A Fast Algorithm of Neural Network for the Training and Recognition of the Phonemes. *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video, and Speech Processing*. 318-321.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustical Characteristics of English Fricatives. *Journal of Acoustical Society of America*. 108: 1252-1263.

Juang, B.H. and Rabiner, L.R. (1991). Hidden Markov Models for Speech Recognition. *Technometrics*. 33(3): 251-271.

Kasuriya, S. , Achariyakulporn, V. , Wutiwiwatchai, C., and Tanprasert, C. (2001). Text-Dependent Speaker Identification via Telephone Based on DTW and MLP. *MIC2001*. 1: 2544-2548.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (2003). Functional Differences between Vowel Onsets and Offsets in Temporal Perception of Speech: Local-Change Detection and Speaking-Rate Discrimination. *Journal of Acoustical Society of America*, 6(113): 3379-3389.

Klatt, D. (1987). Review of Text-to-Speech Conversion for English. *Journal of Acoustical Society of America*. 82: 737-793.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. *Journal of Acoustical Society of America*. 87(2): 820-857.

Koizumi, T., Mori, M., Taniguchi, S., and Maruya, M. (1996). Recurrent Neural Networks for Phoneme Recognition. *Proceeding of International Conference Spoken Language Processing*. 1: 326-329.

Kuah, K., Bodruzzaman, M., and Zein-Sabbato, S. (1994). A Neural Network-Based Text Independent Voice Recognition System. *Proceedings of the IEEE Conference on Creative Technology Transfer – A Global Affair*. 131-135.

Kuhn, M. H., Tomaschewski, H. and Ney, H. (1981 April). Fast Non-Linear Time Alignment for Isolated Word Recognition. *International Conference on Acoustics, Speech, and Signal Processing*. 6: 736-740.

Lawrence, S. and Giles, C.L. (2000). Overfitting and Neural Networks: Conjugate Gradient and Backpropagation. *International Joint Conference on Neural Networks*. 1:114-119.

Lee, C. and Go, J. (1999). Multi-Gradient: A Fast Converging and High Performance Learning Algorithm. *International Joint Conference on Neural Networks*. 3: 1721-1724.

Lee, K-F (1990). Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition. *IEEE Transaction on Acoustics, Speech, and Signal Processing*. 30(4): 599-609.

Lee, T., Ching, P.C., and Chan, L-W. (1998). Isolated Word Recognition using Modular Recurrent Neural Networks. *Journal of Pattern Recognition*. 31(6): 751-760.

Levine, E. (1995). A Time Warping Neural Network. *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*. Detroit, Michigan. 5: 3339-3341.

Li, Y., Rad, A.B., and Peng, W. (1999). An Enhanced Training Algorithm for Multilayer Neural Networks Based on Reference Output Hidden Unit. *Neural Computing and Applications*. 8: 218-225.

Liang, Y. and Liang, X. (2006). Improving Signal Performance of Neural Networks Through Multi-Resolution Learning Approach. *IEEE Transactions on Systems, Man, and Cybernatics – Part B: Cybernatics*. 36(2): 341-352.

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F. S. (1959). Minimal Rules for Synthesizing Speech. *Journal of Acoustical Society of America*. 31: 1490-1499.

Lienard, J. S. and Soong, F. K. (1984 March). On the Use of Transient Information in Speech Recognition. *International Conference on Acoustics, Speech, and Signal Processing.* 9: 9-12.

Lippmann, R. P. (1989). Review of Neural Networks for Speech Recognition. *Neural Computation*. 1: 1-38.

Liu, Y., Lee, Y. C., Chen, H. H., and Sun, G. Z. (1992 June). Speech Recognition using Dynamic Time Warping with Neural Network Trained Templates. *International Joint Conference in Neural Network*. 2: 7-11.

Magimai-Doss, M. (2003). Using Pitch Frequency Information in Speech Recognition. *Proceedings of 8th European on Speech Communication and Technology*. Geneva, Switzerland. 4: 2525-2528.

Mair, S. J. and Shadle, C. H. (1996). The Voiced/Voiceless Distinction in Fricatives: EPG, Acoustic, and Aerodynamic Data. *Proceedings of the Institute of Acoustics*, 18(9): 163-169.

Maragos, P. (1998). Modulation and Fractal Models for Speech Analysis and Recognition. *Proceedings of COST-249 Meeting*. Porto, Portugal. Pg 17.

Mareno, D. M., Jackson, P. J. B., Hernando, J., and Russell, M. J. (2003). Improved ASR in Noise Using Harmonic Decomposition. *International Conference in Phonetic Science*. Barcelona, 1: 14.

Markov, K. and Nakamura, S. (2003). Hybrid HMM/BN LVCSR System Integrating Multiple Acoustic Features. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. I840-I843.

Matsuura, Y, Miyazawa, H., and Skinner, T. E. (1994 September). Word Recognition using a Neural Network and a Phonetically Based DTW. *Proceedings of the 1994 Workshop in Neural Networks for Signal Processing*. 329-334.

McClellan, J. H., Schafer, R. W., and Yoder, M. A. (1998). *DSP First: A Multimedia Approach*. New Jersey: Prentice Hall.

Meyer Eppler, W. (1953). *On the Generating Mechanism of Noise Sounds, Phonetics and Research in Communications*. 7: 196-212. Translated from German text by Michael Hecker.

Mitra, S. K. (2000b). *Digital Signal Processing*, International Editions. Singapore: McGraw Hill.

Muta, H., Baer, T., Wagatsuma, K., Muraoka, T., and Fukada, H. (1988). A Pitch Synchronous Analysis of Hoarseness in Running Speech. *Journal of Acoustical Society of America*. 84(4): 1292-1301.

Myers, C. S. and Rabiner, L. R. (1981). Connected Digit Recognition using a Level-Building DTW Algorithm. *IEEE Transactions on Acoustic, Speech, and Signal Processing.* ASSP-29(3): 351-363.

Morgan, D.P. and Scofield, C.L. (1991). *Neural Networks and Speech Processing*. Boston: Kluwer Academic Publishers.

Narayanan, S. and Alwan, A. (1996). Parametric Hybrid Source Models for Voiced and Voiceless Fricative Consonants. *Proceeding of IEEE-ICASSP*. 1: 377-380.

Narayanan, S. and Alwan, A. (2000). Noise Source Models for Fricative Consonants. *IEEE Transactions on Speech and Audio Processing*. 9(2): 328-344.

Netter, F. H. (1989). *Atlas of Human Anatomy*, Summit, New Jersey: Ciba-Geigy Corporation.

Ng, S. C., Cheung, C. C., Leung, S. H., and Luk, A. (2003). Fast Convergence for Back-Propagation Network with Magnified Gradient Function. *Proceedings of the International Joint Conference on Neural Networks*. 3: 1903-1908.

Nong, T. H., Yunus, J., and Shaikh-Salleh, S. H. (2002). Speaker-Independent Phonation Recognition for Malay Plosives using Neural Networks. *Proceedings of International Joint Conference on Neural Networks.* 1: 619-623.

Nong, T. H., Yunus, J., and Shaikh-Salleh, S. H. (2002). Speaker-Independent Malay Syllable Recognition using Modular Neural Networks. *Proceedings of 2nd World Engineering Congress.* Sarawak, Malaysia. 1: 1-4.

Nong, T. H., Yunus, J., and Wong, L. C. (2002). Speaker-Independent Malay Isolated Sounds Recognition. *Proceedings of the 9th International Conference on Neural Information Processing.* 5: 2405-2408.

O'Shaughnessy, D. (1987). *Speech Communication Human and Machine*. New York: Addison Wesley.

Oglesby, J. and Mason, J. S. (1990). Optimization of Neural Models for Speaker Identification. *Proceedings of International Conference in Acoustics, Speech, and Signal Processing*. 261-264.

Oppenheim, A. V. and Willsky, A. S., and Nawab, S. H. (1997). *Signals and Systems*. 2nd Edition. Upper Saddle River, New Jersey: Prentice Hall.

Parsons, T. W. (1986). *Voice and Speech Processing*. New York : McGraw-Hill.

Patil, P. B. (1998). Multilayered Network for LPC Based Speech Recognition. *IEEE Transactions on Consumer Electronics*. 44(2): 435-438.

Prasanna S R M, Zachariah J M, and Yegnanarayana B (2004). Neural Network Models for Combining Evidence from Spectral and Suprasegmental Features for Text-Dependent Speaker Verification. Proceedings of International Conference on Intelligent, Sensing, and Information Processing. 359-363.

Rabiner, L. and Juang, B. H. (1993). *Fundamentals of Speech Recognition.* Englewood Cliffs, New Jersey: Prentice Hall.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice Hall.

Rothenberg, M. (1968). *The Breath-Stream Dynamics of Simple-Released-Plosive Production*. Switzerland: S. K. Verlag.

Rudasi, L. and Zahorian, S. A. (1991 April). Text-Independent Talker Identification with Neural Networks. *International Conference on Acoustics, Speech, and Signal Processing.* 1: 389-392.

Sae-Tang, S and Tanprasert, C. (May 2000). Feature Windowing for Thai Text-Dependent Speaker Identification using MLP with Back-Propagation Algorithm. *IEEE International Symposium on Circuits and Systems*, Geneva. 3: 579-582.

Sakoe, H. and Chiba, S. (1978 February). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*. ASSP-26(1): 43-49.

Sakoe, H., Isotani, R., and Yoshida, K. (1989). Speaker-Independent Word Recognition using Dynamic Programming Neural Networks. *Proceedings of International Conference in Acoustics, Speech, and Signal Processing.* 1: 29-32.

Salleh, S. H. (1993). A Comparative Study of the Traditional Classifier and the Connectionist Model for Speaker Dependent Speech Recognition System. Universiti Teknologi Malaysia: Master Thesis.

Salleh, S. H. (1997). *An Evaluation of Preprocessors for Neural Network Speaker Verification*. University of Edinburgh, UK: Ph.D. Thesis.

Sarkar, D. (1995). Methods to Speed Up Error Back-Propagation Learning Algorithm. *ACM Computing Surveys*. 27(4): 519-542.

Scheper, R. A. and Teolis, A. (2003). Cramer Rao Bounds for Wavelet Transform Based Instantaneous Frequency Estimates. *IEEE Transactions on Signal Processing*. 51(6): 1593-1603.

Scully, C., Castelli, E., Brearley, E., and Shirt, M. (1992). Analysis and Simulation of Speaker's Aerodynamic and Acoustic Patterns for Fricatives. *Journal of Acoustical Society of America*. 20: 39-51.

Sedik, H., Rahmouni, A, and Sayadi, M. (2004). Text Independent Speaker Recognition using the Mel Frequency Cepstral Coefficients and a Neural Network Classifier. *International Symposium on Control, Communications, and Signal Processing*. Tunis, Tunisia, 631-634.

Shadle, C. H. (1995). Modeling the Noise Source in Voiced Fricatives. *Proceedings of the National Congress on Acoustics*. Trodheim, Germany, 3: 145-148.

Shadle, C. H. and Mair, S. J. (1996). Quantifying Spectral Characteristics of Fricatives. *Proceeding of ICSLP*. Philadelphia, 1521-1524.

Sima, M., Croitoru, V., and Burileanu, D. (1998). Performance Analysis on Speech Recognition using Neural Networks. Proceeding of the Device and Application System. Suceava, Romania. 259-266.

Singer, H. And Sagayama, S. (1992). Pitch Dependent Phone Modelling for HMM Based Speech Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1: 273-276.

Soens, P. and Verhelst, W. (2005). Split Time Warping for Improved Automatic Time Synchronization of Speech. *Proceeding of SPS DARTS*, Antwerp, Belgium.

Stephenson, T.A., Escofet, J., Magimai-Doss, M., and Bourlard, H. (2002). Dynamic Bayesian Network Bases Speech Recognition with Pitch and Energy as Auxiliary Variables. *12th IEEE Workshop on Neural Networks for Signal Processing*. 637-646.

Stephenson, T.A., Magimai-Doss, M., and Bourlard, H. (2004). Speech Recognition with Auxiliary Information. *IEEE Transactions on Speech and Audio Processing*. 12(3): 189-203.

Stevens, K. N. (1971). Airflow and Turbulence Noise for Fricatives and Stop Consonants: Static Considerations. *Journal of Acoustical Society of America*. 51: 1180-1192.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., and Kurowski, K. (1992). Acoustic and Perceptual Characteristics of Voicing in Fricatives and Fricative Clusters. *Journal of Acoustical Society of America*. 91(5): 2979-3000.

Sudirman, R., Salleh, Sh-H., and Ming, T. C. (2005a). Pre-Processing of Input Features using LPC and Warping Process. *Proceeding of International Conference on Computers, Communications, and Signal Processing*. 300-303.

Sudirman, R., Salleh, Sh-H., and Ming, T.C. (2005b). NN Speech Recognition Utilizing Aligned DTW Local Distance Scores. *Proceeding of 9th International Conference on Mechatronics Technology*. ICMT-192.

Sudirman, R., Salleh, Sh-H., Khalid, P. I., Ahmad, A. H. (2005c). Normalization of LPC Feature using Warping Method. *Jurnal Elektrika*. 2(7): 29-35.

Sudirman, R., Salleh, Sh-H., and Salleh, S. (2006). Local DTW Coefficients and Pitch Feature for Back-Propagation NN Digits Recognition. *Proceedings of IASTED International Conference on Networks and Communications System*. 201-206.

Sun, F., Li, B., and Chi, H. (1991). Some Key Factors in Speaker Recognition using Neural Networks Approach. *IEEE International Conference on Neural Networks*. Singapore. 3: 2725-2756.

Swee, T. T. (2003). *Malay Text to Speech*. Universiti Teknologi Malaysia: Master Thesis.

Sze, H. K. (2004). *The Design and Development of an Educational Software on Automatic Speech Recognition*. Universiti Teknologi Malaysia: Master Thesis.

Tabatabaee, V., Azimisadjadi, B., Zahirazami, S. B., and Lucas, C. (1994). Isolated Word Recognition using a Hybrid Neural Network. *Proceedings of International Conference in Acoustics, Speech, and Signal Processing*. 2(2): 649-652.

Tebelskis, J, Waibel, A, Petek, B., and Schmidbauer, O. (1991 April). Continuous Speech Recognition using Linked Predictive Neural Networks. *International Conference on Acoustics, Speech, and Signal Processing.* 1: 61-64.

Titze, I. R. (1994). *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice Hall.

Tritton, D. J. (1988). *Physical Fluid Dynamics*. 2nd Edition. New York, NY: Oxford University Press.

Tsai, H. L. and Lee, S. J. (1997 October). A Neural Network Model for Spoken Word Recognition. *IEEE International Conference on Systems, Man, and Cybernetics*. 5: 4029-4034.

Uma, S., Sridhar, V., and Krishna, G. (Sept. 1992). Time-Normalization Techniques for Speaker-Independent Isolated Word Recognition. *Proceedings of Pattern Recognition Conference: Image, Speech and Signal Analysis*. 3: 537-540.

Unal, F.A. and Tepedelenlioglu, N. (1992). Dynamic Time Warping Using an Artificial Neural Networks. *International Joint Conference on Neural Networks*. 4: 715 – 721.

Welling, L. and Ney, H. (1998). Formant Estimation for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*. 6(1): 36-48.

Wildermoth, B. R. (2001). *Text-Independent Speaker Recognition using Source Based Features*. Griffith University, Australia: Master of Philosophy Thesis.

Wong, P-F. and Siu, M-H. (2002). Integration of Tone Related Feature for Chinese Speech Recognition. *6th International Conference on Signal Processing*. 1: 476-479.

Woodland, P.C., Odell, J.J., Valtchec, V., and Young, S.J. (1994). Large Vocabulary Continuous Speech Recognition using HTK. *International Conference on Acoustics, Speech, and Signal Processing*. 1-4.

Wouhaybi, R. H. and Al-Aloui, M. A. (1999). Comparison of Neural Networks for Speaker Recognition. *Proceedings of The Sixth IEEE International Conference on Electronics, Circuits and Systems*. Pafos, Cyprus, 1: 125-128.

Wu, J. and Chan, Chorkin (1993). Isolated Word Recognition by Neural Network Models with Cross-Correlation Coefficients for Speech Dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 15(11): 1174-1185.

Yu, X.H., Chen, G.A., and Cheng, S.X. (1995). Dynamic Learning Rate Optimization of the Back-Propagation Algorithm. *IEEE Transactions on Neural Networks*. 6(3): 669-677.

Zbancioc, M and Costin, M. (2003). Using Neural Networks and LPCC to Improve Speech Recognition. *International Symposium on Signals, Circuits, and Systems*. 2: 445-448.

Zhao, Y. (1993). A Speaker-Independent Continuous Speech Recognition System using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units. *IEEE Transactions on Speech and Audio Processing*. 1(3): 345-361.

Zhu, M. and Fellbaum, K. (1990 April). A Connectionist Model for Speaker-Independent Isolated Word Recognition. *International Conference on Acoustics, Speech, and Signal Processing.* 1: 529-532.