ISOLATED ENGLISH ALPHABET SPEECH RECOGNITION USING WAVELET CEPSTRAL COEFFICIENTS AND NEURAL NETWORK

TARMIZI ADAM

UNIVERSITI TEKNOLOGI MALAYSIA

Replace this page with form PSZ 19:16 (Pind. 1/07), which can be obtained from SPS or your faculty.

ISOLATED ENGLISH ALPHABET SPEECH RECOGNITION USING WAVELET CEPSTRAL COEFFICIENTS AND NEURAL NETWORK

TARMIZI ADAM

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

Faculty of Computing
Universiti Teknologi Malaysia

MARCH 2014

I dedicate this thesis to my lovely parents...

ACKNOWLEDGEMENT

Alhamdulillah, Praise be to Allah the Almighty. Without His help this thesis would not have been possible. Whoever He guides non can misguide, whomsoever He misguide non can guide. My first thank you goes to my supervisor Dr Md Sah Hj Salam for his professionalism and enthusiastic guidance during my two years at University Teknologi Malaysia (UTM). His style of flexibility and freedom ensured students to be more creative.

I would also like to take this opportunity to thank my co-supervisor at International Islamic University Malaysia (IIUM) Dr Teddy Surya Gunawan for his guidance in the field of wavelets and signal processing. His help especially in MATLAB programming also proved to be invaluable to me especially to write better code for this study.

Next I would like to thank my lab mates for their advices and insights over the past two years of my stay at UTM. For this, special credits go to Khalid for introducing me to Latex. Without him introducing me to Latex I would have suffered greatly to produce this thesis maybe spending more of my time formating rather than writing the contents of the thesis itself. Brother Isma also deserves special thanks for his insights and discussion. To Brother Ku Faisal, I would also like to thank him for his discussion on outdoor activities that we went together which made my stay at UTM more memorable.

A special thanks also goes to brother Suhaimi. He was the one who familiarized my stay at UTM and various places at Skudai and Johor Baharu. Our discussion on various topics about research and everyday life benefited me directly and indirectly. His warm and friendly attitude made my stay at UTM meaningful.

Finally, I would like to thank my beloved parents Adam Puteh and Faizah Mustaffa for their constant support and love and for teaching me that the best

investment one can have in life is education.

ABSTRACT

Speech recognition has many applications in various fields. One of the most important phase in speech recognition is feature extraction. In feature extraction relevant important information from the speech signal are extracted. However, two important issues that affect feature extraction are noise robustness and high feature dimension. Existing feature extraction which uses fixed windows processing and spectral analysis methods like Mel-Frequency Cepstral Coefficient (MFCC) could not cater robustness and high feature dimension problems. This research proposes the usage of Discrete Wavelet Transform (DWT) to replace Discrete Fourier Transform (DFT) for calculating the cepstrum coefficients to produce a newly proposed Wavelet Cepstral Coefficient Wavelet Cepstral Coefficient (WCC). The DWT is used in order to gain the advantages of the wavelet in analyzing non stationary signals. The WCC is computed in a frame by frame manner. Each speech frame is decomposed using the DWT and the log energy of its coefficients is taken. The final stage of the WCC computation is done by taking the Discrete Cosine Transform (DCT) of these log energies to form the WCC. The WCC are then fed into a Neural Network (NN) for classification. In order to test the proposed WCC a series of experiments were conducted on TI-ALPHA dataset to compare its performance with the MFCC. The experiments were conducted under several noise levels using Additive White Gaussian Noise (AWGN) and number of coefficients for speaker dependent and independent tasks. From the results, it is shown that the WCC has the advantage of withstanding noisy conditions better than MFCC especially under small number of features for both speaker dependent and independent tasks. The best result tested under noisy condition of 25 dB shows that 30 WCC coefficients using Daubechies 12 achieved 71.79% recognition rate in comparison to only 37.62% using MFCC under the same constraint. The main contribution of this research is the development of the WCC features which performs better than the MFCC under noisy signals and reduced number of feature coefficients.

ABSTRAK

Pengecaman suara mempunyai pelbagai aplikasi dalam berbagai bidang. Salah satu fasa yang terpenting bagi pengecaman suara ialah penyarian ciri. Pada fasa penyarian ciri informasi penting pada isyarat bunyi disari. Walaubagaimanapun, dua isu penting yang mempengaruhi penyarian ciri adalah keteguhan pada hingar dan jumlah ciri yang besar. Teknik-teknik penyarian ciri yang sedia ada seperti Pekali Cepstral Frekuensi Mel (MFCC) memproses isyarat suara dengan menggunakan bingkai bersaiz tetap dan menggunakan analisis spektral tidak mampu menanggani masalah keteguhan pada hingar dan jumlah ciri yang besar. Kajian ini mencadangkan penggunaan Transformasian Wavelet Diskret (DWT) bagi menggantikan Transformasian Fourier Diskret (DFT) untuk mengira pekali cepstrum bagi menghasilkan ciri baru yang dipanggil Pekali Wavelet Cepstral (WCC). Penggantian menggunakan DWT adalah disebabkan kelebihan yang terdapat pada wavelet dalam menganalisa isyarat pegun. Pengiraan WCC dilaksanakan pada setiap bingkai isyarat suara. Setiap bingkai isyarat suara diurai menggunakan DWT dan tenaga logaritma pekalinya diambil. Langkah terakhir dalam pengiraan WCC dibuat dengan mengira Transformasian Kosinus Diskret (DCT) tenaga logaritma tersebut bagi menghasilkan WCC. Ciri WCC ini kemudianya disuap ke Rangkaian Neural (NN) bagi tujuan kalsifikasi. Bagi menguji ciri baru WCC yang dicadangkan, beberapa siri eksperimen telah dijalankan pada data suara TI-ALPHA bagi tujuan perbandingan prestasi dengan ciri MFCC. Ujian telah dilakukan dengan mengambil kira beberapa tingkatan hingar menggunakan Hingar Putih Gaussian (AWGN) dan saiz ciri untuk pengecaman kebergantungan pengucap dan tidak kebergantungan pengucap. Keputusan terbaik pada kondisi hingar 25 dB menunjukkan 30 pekali WCC menggunakan Daubechies 12 memperoleh pengecaman sebanyak 71.97% dibandingkan dengan hanya 37.62% menggunakan MFCC pada kekangan yang sama. Sumbangan utama kajian ini adalah menghasilkan ciri WCC yang mempunyai prestasi pengecaman yang lebih baik dari ciri MFCC pada hingar yang tinggi dan jumlah ciri yang kecil.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEMENT	iv
	ABSTRACT	vi
	ABSTRAK	vii
	TABLE OF CONTENTS	viii
	LIST OF TABLES	xii
	LIST OF FIGURES	xiv
	LIST OF APPENDICES	xvi
	LIST OF ABBREVIATIONS	xvii
1	**INTRODUCTION**	**1**
	1.1 Introduction	1
	1.2 Problem Background	3
	1.3 Problem Statement	4
	1.4 Research Aims	6
	1.5 Objectives	6
	1.6 Research Scope	6
	1.7 Importance of Study	7
	1.8 Thesis Overview	7
2	**LITERATURE REVIEW**	**9**
	2.1 Introduction	9
	2.2 Automatic Speech Recognition	10
	2.3 General Framework of the Automatic Speech Recognition (ASR)	13
	2.3.1 Pre-Processing	13
	2.3.2 Feature Extraction	15

		2.3.3	Classification	16
	2.4	Artificial Neural Networks		16
		2.4.1	Multilayer Perceptron and Back Propagation Algorithm	17
	2.5	Neural Network Parameters		20
	2.6	Fundamentals of Cepstral Analysis		21
		2.6.1	Computing the Cesptral Coefficients	22
	2.7	Mel-Frequency Cepstral Coefficient (MFCC)		23
	2.8	Wavelets		26
		2.8.1	Discrete Wavelet Transform	28
		2.8.2	Wavelet Feature Extraction	29
		2.8.3	Hybrid Wavelet Feature Extraction	32
		2.8.4	Reducing Feature Dimension	34
	2.9	Noise in Speech Signals		38
	2.10	Speech Dataset		40
	2.11	Chapter Summary		41
3	**RESEARCH METHODOLOGY**			**42**
	3.1	Introduction		42
	3.2	Research Framework		42
		3.2.1	Literature Review	43
		3.2.2	Problem Formulation	44
		3.2.3	Data Collection	44
		3.2.4	Pre-processing	44
		3.2.5	Development of the Wavelet Cepstral Coefficients	45
		3.2.6	Development of Speech Recognition System	45
		3.2.7	Evaluation and Result	46
	3.3	Summary		46
4	**NEURAL NETWORK OPTIMIZATION**			**47**
	4.1	Introduction		47
	4.2	Feature Extraction for Training and Testing		48
		4.2.1	Training	48
		4.2.2	Testing	49
	4.3	Setting the Neural Network		50
		4.3.1	Finding a Fixed Number for Input Nodes	50

		4.3.2	Finding the Number of Nodes in the Hidden Layer	53
		4.3.3	Learning Rate and Momentum Constant	53
	4.4	Obtaining the Best learning rate and momentum constant		54
		4.4.1	Suggested learning rate and momentum constant	54
		4.4.2	Finding the learning rate and momentum constant	55
	4.5	Designing the Neural Network Targets		57
	4.6	Normalizing the inputs to the Neural Network		58
	4.7	Chapter Summary		59
5	**SPEECH RECOGNITION USING THE MEL-FREQUENCY CEPSTRAL COEFFICIENTS**			**60**
	5.1	Speech Recognition Under Clean Speech		60
		5.1.1	Experimental Setup	60
		5.1.2	Result for Speech Recogniton Under Clean Speech	61
		5.1.3	Discussion	63
	5.2	Speech Recognition Under Noisy Environments		63
		5.2.1	Experimental Setup	64
		5.2.2	Results	64
			5.2.2.1 Noisy Training and Testing Features	64
			5.2.2.2 Noisy Training Features with Clean Testing Features	66
			5.2.2.3 Clean Training Features With Noisy Testing Features	68
		5.2.3	Discussion	69
			5.2.3.1 Noisy Training and Testing Features	69
			5.2.3.2 Noisy Training Features with Clean Testing Features	70
			5.2.3.3 Clean Training Features with Noisy Testing Features	71
	5.3	Speech Recognition Under Reduced Feature Dimension		72
		5.3.1	Experimental Setup	73

		5.3.2 Results	74
		5.3.3 Discussion	75
6	**SPEECH RECOGNITION USING THE WAVELET CEPSTRAL COEFFICIENTS**		**77**
	6.1	Introduction	77
	6.2	Wavelet Cepstral Coefficients	77
		6.2.1 Technical Background	78
	6.3	Speech Recognition Under Clean Speech	82
		6.3.1 Experimental Setup	82
		6.3.2 Results	83
		6.3.3 Discussion	84
	6.4	Speech Recognition Under Noisy Speech	86
		6.4.1 Experimental Setup	86
		6.4.2 Results	88
		6.4.3 Discussion	88
7	**CONCLUSION AND FUTURE WORKS**		**96**
	7.1	Summary	96
	7.2	Contributions and Findings	97
	7.3	Research limitations	97
	7.4	Future Work	98
REFERENCES			**99**

LIST OF TABLES

TABLE NO.	TITLE	PAGE
1.1	Problems with MFCCs	5
2.1	Phases of Automatic Speech Recognition (ASR) development	12
2.2	Advantages of Mel-Frequency Cepstral Coefficient (MFCC)	23
2.3	Justification of using wavelets as a feature extractor	35
2.4	Some comments regarding previous research	36
2.5	Speech dataset summary	40
4.1	feature extraction step for training.	48
4.2	MFCC parameter settings	49
4.3	Summary of training and testing data	49
4.4	NN classifier setup.	50
4.5	Steps for estimating a fixed feature vector size.	52
4.6	Learning rate and momentum constant used for initial experimentation.	54
4.7	Several learning rate and momentum constant pairs as experimented and suggested in (Salam *et al.*, 2009, 2011).	55
4.8	Best learning rate and momentum constant obtained.	56
4.9	Best learning rate and momentum constant.	57
5.1	MFCC parameter settings	61
5.3	Effects of noise for speaker independent recognition	65
5.2	Effects of noise for speaker dependent recognition	65
5.4	Effects of noise for speaker dependent recognition	66
5.5	Effects of noise for speaker independent recognition	67
5.6	Effects of noise for speaker dependent recognition	68
5.7	Effects of noise for speaker independent recognition	68
5.8	Steps for MFCC feature truncation using zero-padded normalization.	72

5.9	Effects of reduced feature dimension for speaker dependent task	74
5.10	Effects of reduced feature dimension for speaker independent task	74
6.1	Settings for the WCC experiment for clean speech	83
6.3	WCC performance with different feature dimension for speaker independent task	84
6.2	WCC performance with different feature dimension for speaker dependent task	84
6.4	Settings for the WCC experiment for noisy speech	87
6.5	Settings for the MFCC experiment for noisy speech	87
A.1	Results for speaker dependent and independent tasks using Daubechies wavelets (40, 60, and 90 coefficients)	103
A.2	Results for speaker dependent and independent tasks using Daubechies wavelets (120, 180, and 270 coefficients)	104
A.3	Results for speaker dependent and independent tasks using Coiflet wavelets (120, 180, and 270 coefficients)	104

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
1.1	Speech recognition framework	2
2.1	Parts of an analog-to-digital converter	15
2.2	Two neuron models. (a) A biological neural network. (b) Perceptron model by McCulloch. Figures after (Negnevitsky, 2005)	18
2.3	Three layer back-propagation neural network (Negnevitsky, 2005).	19
2.4	Common steps for the MFCC feature extraction process.	24
2.5	Example of two different wavelet functions. (a) Morlet wavelet. (b) Mexican hat wavelet.	27
2.7	Mismatch as a result of noisy conditions (Xiao, 2009)	39
2.6	Additive noise	39
3.1	Research framework	43
4.1	The sub figures are: (a) utterance of the letter ’R’ from female speaker 4 (F4); (b) utterance of the letter ’R’ from male speaker 6 (M6).	51
4.2	Average recognition rate of some suggested learning rate and momentum constant.	55
4.3	Average recognition rate for learning rate and momentum constant in Table 4.8.	56
4.4	Illustration of the target matrix	58
5.1	Recognition rates. (a) Speaker dependent. (b) Speaker independent.	62
5.2	Effects of noise for speaker dependent and speaker independent recognition	66
5.3	Effects of noise for speaker dependent and speaker independent recognition	67
5.4	Effects of noise for speaker dependent and speaker independent recognition	69

5.5	Temporal normalization methods. (a) Zero padding. (b) Truncation.	73
5.6	Average recognition rates for both speaker dependent and speaker independent under different number of MFCC coefficients.	75
6.1	Block diagram of WCC	78
6.2	Overall computation steps of the WCC with 3 levels of DWT. The WCC is computed in a frame by frame manner.	81
6.3	MFCC and WCC comparisons under clean speech with different coefficient sizes. (a) Speaker dependent. (b) Speaker independent.	85
6.4	Three experimental setup results for 30 coefficient WCC under various noise levels.	89
6.5	Three experimental setup results for 60 coefficient WCC under various noise levels.	90
6.6	Three experimental setup results for 90 coefficient WCC under various noise levels.	91
6.7	Three experimental setup results for 120 coefficient WCC under various noise levels.	92
6.8	Three experimental setup results for 180 coefficient WCC under various noise levels.	93
6.9	Example of a gradient error surface	94
6.10	Graphical explanation of the different results for the three experimental setup environments. (a) and (b) Experiment 1. (c) and (d) Experiment 2. (e) and (f) experiment 3. Here, depending on the experiment, the training or testing are either clean or at 0 dB.	95

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	PRELIMINARY EXPERIMENTAL RESULTS	103
B	PULBLICATIONS	106

LIST OF ABBREVIATIONS

AFLPC	–	Average Framing Linear Predictive Coding
AI	–	Artificial Intelligence
ANN	–	Artificial Neural Network
ASR	–	Automatic Speech Recognition
AWGN	–	Additive White Gaussian Noise
CMU	–	Carnegie Mellon University
DARPA	–	Defense Advanced Research Project Agency
DCT	–	Discrete Cosine Transform
DFT	–	Discrete Fourier Transform
DTW	–	Dynamic Time Warping
DWPT	–	Discrete Wavelet Packet Transform
DWT	–	Discrete Wavelet Transform
DWLPC	–	Dyadic Wavelet Decomposition Linear Predictive Coefficient
FFT	–	Fast Fourier Transform
GWN	–	Gaussian White Noise
HLDA	–	Heteroscedastic Linear Discriminant Analysis
HMM	–	Hidden Markov Model
IBM	–	International Business Machine
LDA	–	Linear Discriminant Analysis
LELVM	–	Laplacian Eigenmaps Latent Variable Model
LPC	–	Linear Predictive Coefficient
LPCC	–	Linear Predictive Cepstral Coefficient
MFCC	–	Mel-Frequency Cepstral Coefficient
MFDWC	–	Mel-Frequency Discrete Wavelet Coefficients
MLP	–	Multilayer Perceptron
MSE	–	Mean Square Error
NN	–	Neural Network

PCA	–	Principle Component Analysis
PLP	–	Perceptual Linear Prediction
PR	–	Pattern Recognition
RASTA	–	Relative Spectra
SNR	–	Signal to Noise Ratio
SBC	–	Subband Based Cepstral
STFT	–	Short Time Fourier Transform
TI	–	Texas Instruments
TIMIT	–	Texas Instrument Massachusetts Institute of Technology
UWLPC	–	Uniform Wavelet Decomposed Linear Predictive Coefficient
WCC	–	Wavelet Cepstral Coefficient
WP	–	Wavelet Packet
WPCC	–	Wavelet Packet Cepstral Coefficients
WPP	–	Wavelet Packet Parameters

CHAPTER 1

INTRODUCTION

1.1 Introduction

Speech recognition over the past few decade has been an emerging field thanks to the advance in computational power of computers and ongoing research, development and discoveries in the field of speech processing, audio and acoustic. These discoveries and breakthroughs have helped the field of speech recognition mature over time. Speech recognition is in fact an interesting field combining various other fields such as computer science, engineering, linguistics and security. In fact, it is an interesting field of human computer interaction. Speech recognition may enable humans to interact with machines more naturally as speech is one of the most natural form of human interaction.

The framework of speech recognition system has several steps as shown in Figure 1.1. These steps are divided into pre-processing, Feature extraction and classification and is identical in almost all practical pattern recognition system (Pandya and Macy, 1996). Each of these steps plays an important role in order for a speech recognition system to function accurately and reliably. The first part which is the pre-processing stage is usually concerned with speech processing such as analog to digital conversion of speech, and speech enhancement techniques. Feature extraction which is the second part of the framework deals with extracting certain unique features from the signal that may contain significant information. In speech recognition systems, features extracted must be unique to a particular word or utterances in order to aid the classification step.

The final step in the speech recognition framework is the classification step. As the name implies, this step classifies or recognize the utterance or speech fed by the user of the speech recognition system. This step heavily employs Pattern Recognition (PR) and Artificial Intelligence (AI) techniques as its main driving force.Although the pre-processing and the classification part are vital components of any speech recognition systems, the feature extraction plays a very important role in the accuracy of speech recognition systems. In fact, it is also a very important step in almost all PR

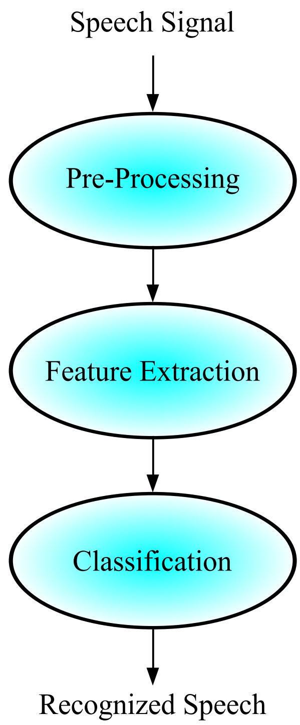

Figure 1.1: Speech recognition framework

processes. Feature extracted must be robust to corrupted, degraded and noisy speech signal i.e. even when the speech signal is subjected to various interferences, good features may still be extracted and used for classification to yield accurate and reliably recognition accuracy.

Another important aspect for the feature extracted is their ability to store unique information regarding confusable acoustic feature of speech. For example the acoustic similarity between letters 'B' and 'D' are in fact very hard to discriminate. Hence, feature extracted from these confusable acoustic speeches must have the ability to precisely store relevant acoustical features that will help the classification step recognize the input speech.

Traditional speech feature like the Linear Predictive Coefficient (LPC), Linear Predictive Cepstral Coefficient (LPCC), and the MFCC shows high performance when used under benign conditions however, their performance decreases under the influence of background noise and degradation which is particularly true for MFCCs (Wu and Lin, 2009). A pure use of MFCCs as signal features has also shown to exhibit lower performance in dealing with confusable acoustic sets compared to other refined techniques as shown by Karnjanadecha and Zahorian (2001)

In order to address the problem regarding the weakness of the MFCCs, this research focuses on the feature extraction phase within the speech recognition framework. An improvement for the cepstral coefficients to withstand degraded speech is proposed. The proposed technique will be tested and evaluated with isolated English alphabets. Theories, algorithms, simulations and results will be provided to prove the proposed method in this study.

1.2 Problem Background

The need for accurate and reliable speech recognition system for application in security systems, telephony and dictation poses a great challenge in the field of ASR. One of the challenges is to extract speech signal features that can best represent or discriminate among classes. Because of various interferences introduce to the speech signal prior to the feature extraction phase, features extracted from the signals are contaminated and may lead to accuracy reduction. This is somewhat true in almost all practical environments. Thus, a need for speech feature that can accurately classify and discriminate speech under practical environment that consist background noise and degradation must be addressed. This would better aid the pre-processing and classification stage and hence, yield better recognition rates for speech recognition systems.

Conventional ASR system uses several feature extraction techniques to extract distinct features from an input speech signal. However, the most popular and widely used among feature extraction technique is the MFCC (Gowdy and Tufekci, 2000; Wu and Lin, 2009). MFCCs are used because it models the human auditory perception with regard to frequencies which in return can represent sound better (Abdallah and Ali, 2010; Razak *et al.*, 2008).

Even though MFCCs are widely used and has its own advantages, the problem arises when the input speech signal is not clean or degraded. As a result, the MFCC features extracted from these signal could be said as ”contaminated” as these features also incorporate the distortions or degradations that were present in the signal. Thus, MFCCs have a poor ability to withstand noise and degradations (Anusuya and Katti, 2011; Sarikaya *et al.*, 1998).

Another problem with the MFCC feature is that it assumes each speech frame used in its computation to be fixed in length and the signal analyzed within it as stationary which in practice, is not quite true (Daqrouq and Al Azzawi, 2012; Shafik *et al.*, 2009). With fixed analyzing frames, localized events and abrupt changes in the speech signal are poorly analyzed. These localized events of speech may contain important information that can affect the recognition rate of a speech recognizer (Gowdy and Tufekci, 2000).

From a classification perspective, the MFCC extracted from a signal produces large number of features (Jafari and Almasganj, 2010). For example, each frame of speech produces 13 static MFCC coefficients while 39 coefficients for each frame when the dynamic features (velocity and acceleration) are also extracted. The large feature inputed to the classifier will require a computational expensive recognizer (Flynn and Jones, 2012a; Paliwal, 1992). This is particularly true in NN speech recognition systems. In a NN speech recognizer the number if input depends exactly on the number of features extracted from the speech signal. Thus, a large number of feature requires a large number of input nodes resulting in a computational extensive recognizer. Furthermore, large number of features also requires large number of storage.

MFCC features used in NN speech recognition system often uses high number of input nodes because of the large number of features extracted. For example Salam *et al.* (2009, 2011) requires 820 inputs for connected Malay digit and 360 inputs for isolated Malay digits. While, for English alphabets Cole and Fanty (1990) used 617 inputs to the NN classifier. Reducing the number of feature used is also an important aspect in a distributed speech recognition system in where feature extraction and classification are done separately between a client and server (Flynn and Jones, 2012a). Thus the problem of reducing the number of features used while acquiring good recognition rate is another issue to consider. Table 1.1 summarizes several problems of the MFCCs.

In order to address these problems, many researches were conducted to further enhance the capability of the MFCCs. One such way used by various researchers was to adopt the wavelet transform. Research have shown that wavelet transform are robust to noise and degradations (Farooq and Datta, 2004; Flynn and Jones, 2012b; Gowdy and Tufekci, 2000). Combination of wavelets with other feature extraction techniques as a hybrid feature extractor have also yielded better recognition rates (Abdallah and Ali, 2010; Al-Sawalmeh *et al.*, 2010; Shafik *et al.*, 2009; Zhang *et al.*, 2006). For feature reduction, recent studies show their effectiveness in producing small feature dimension as shown by Flynn and Jones (2012b).

1.3 Problem Statement

Mining the problem of the cepstral analysis methods which uses the DFT where the analysis of speech signal is done in a fixed window setting and the issue of large feature dimension produced especially from MFCC feature extraction we propose the use of wavelets for cepstral analysis rather than the DFT. By using wavelets, the speech signal is analyzed with a non fixed window scheme. With non fixed window analysis, high frequency regions of the signal are analyzed with small windows while low frequency regions are analyzed with large windows. With this, more local information

Table 1.1: Problems with MFCCs

Problem	**Description**
Robustness issues	As stated by Anusuya and Katti (2011); Sarikaya *et al.* (1998) MFCCs are not immune to noise as an example in telephone speech where the speech is degraded by convolutional channel noise. MFCCs are easily corrupted as it uses DCT of the mel-scaled log filterbank energy. DCT covers all frequency bands thus a corruption in a frequency band effect the whole of MFCCs(Gowdy and Tufekci, 2000)
Fixed window/frame length	As pointed out by Gowdy and Tufekci (2000), MFCCs uses fixed window or frame of speech which means that it assumes that only one information at a time is conveyed. This is not true as some frame might have voiced and unvoiced sounds simultaneously
Large feature numbers	From a recognizers perspective especially NN based speech recognizers, the large numbers of features extracted effects the computational cost of the recognizer (Flynn and Jones, 2012a). For NN recognizers the number of input to the NN are highly dependent on the number of features extracted.

such as transients are extracted. Moreover, by using wavelets the size of the feature vector can be effectively reduced. Thus, the primary research question for this research is stated as:

- ***Can combination of DWT and cepstral analysis feature improve recognition rate under noisy data and reduced feature dimension using neural networks?***

To address the research questions posed, the research aims and objectives have been identified and will be presented in the next section.

1.4 Research Aims

The aim of this research is to propose a new feature extraction method by using the DWT to compute the cepstrum of a speech signal therefore, a new set of speech feature called WCC will be derived.

1.5 Objectives

In order to address the problem and achieve the aim of this study, the objectives of the research are:

1. To develop a new speech feature by the use of DWT and cepstrum.
2. To test the developed features with English Isolated speech database in benign and noisy conditions using neural network.
3. Compare the proposed features with MFCC.

1.6 Research Scope

To narrow the research scope and to be parallel with the problems and research objectives the following scope has been agreed upon;

1. This research will be focusing on the feature extraction process within the SR framework (Refer Figure 1.1).
2. Recognition task will be Isolated words.
3. Isolated words that will be used are from the English Texas Instruments (TI) 46 dataset alphabets.
4. For simulating speech signal degradation, AWGN will be used.

1.7 Importance of Study

Speech recognition systems has many limitations especially in adverse or noisy environments. In this thesis, we propose a new feature which has a small feature dimension for neural network speech recognizer and invariant to noise. The contribution of this thesis include:

1. A new set of features called WCC which are more noise invariant under small feature dimensions. This is particular important in NN speech recognizers where the input nodes are directly proportional to the number of speech features. Low input nodes are desirable to decrease computational complexity.
2. An evaluation of NN based speech recognizer under various noisy conditions and the effects of the NN learning rate and momentum constant under these conditions. The results shows the importance of choosing suitable learning rate and momentum constant for a specific task.

1.8 Thesis Overview

The thesis is organized as follows. Chapter 2 reviews some of the technical background that are related and will be used in this thesis.

Chapter 3 presents the NN setup especially in estimating the most suitable learning rate and momentum constant that will be used for almost all clean speech experiments. Several initial speech recognition experiments using various values of learning rate and momentum constant were conducted for this purpose.

In chapter 4, we proceed with the MFCC experiments. Here, raw MFCC features are used for recognizing 26 English alphabets. Various noise level are tested with three different training and testing environments. The result in this chapter will be used for benchmarking purpose for our proposed WCC features.

In chapter 5 we conduct our proposed WCC experiments and compare the results with the MFCCs. Similar to the experiments in chapter 4 we conduct the experiments for the WCCs in various noise level with three different training and testing environments. We also explore the ability of the WCC withstand feature dimension while preserving higher recognition then the MFCCs. This is done by varying the number of WCC coefficients to the NN classifier.

Chapter 6 concludes this thesis and provides some contribution and suggestion for further works.

REFERENCES

Abdallah, M. I. and Ali, H. S. (2010). Wavelet Based Mel Frequency Cepstral Coefficients for Speaker Identification using Hidden Markov Models. *Journal of Telecommunications*. 1(2), 16–21.

Abushariah, M. A.-A. R. M. (2006). *A Vector Quantization Approach to Isolated Word Automatic Speech Recognition*. Thesis.

Al-Haddad, S. A. R., Samad, S. A., Hussain, A. and Ishak, K. A. (2008). Isolated Malay Digit Recognition using Pattern Recognition Fusion of Dynamic Time Warping and Hidden Markov Model. *American Journal of Applied Science*. 5, 714–720.

Al-Sawalmeh, W., Daqrouq, K., Daoud, O. and Al-Qawasmi, A. R. (2010). Speaker Identification System-Based Mel Frequency and Wavelet Transform using Neural Network Classifier. *European Journal of Scientific Research*. 41, 515–525.

Alotaibi, Y. (2009). A Simple Time Alignment Algorithm for Spoken Arabic Digit Recognition. *Journal of King Abdulaziz University: Engineering Sciences*. 20(1), 29–43.

Anusuya, M. and Katti, S. (2011). Front End Analysis of Speech Recognition: A Review. *International Journal of Speech Technology*. 14(2), 99–145.

Ayadi, M. E., S.Kamel, M. and Karray, F. (2011). Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition*. 44(3), 572–587.

Coifman, R. R. and Wickerhauser, M. V. (1992). Entropy-Based Algorithms for Best Basis Selection. *IEEE Transactions on Information Theory*. 38(2), 713–718.

Cole, R. and Fanty, M. (1990). Spoken letter recognition. In *Proceedings of the workshop on Speech and Natural Language*. HLT '90. 385–390.

Daqrouq, K. (2011). Wavelet Entropy and Neural Network for Text-Independent Speaker Identification. *Engineering Applications of Artificial Intelligence*. 24(5), 796–802.

Daqrouq, K. and Al Azzawi, K. Y. (2012). Average Framing Linear Prediction Coding with Wavelet Transform for Text-Independent Speaker Identification System. *Computers and Electrical Engineering*. (0).

Davis, S. and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 28(4), 357–366.

Deller, J. R., Hansen, J. H. and Proakis, J. G. (2000). *Discrete Time Processing of Speech Signals*. New York: Wiley.

Deng, L. and Huang, X. (2004). Challenges in Adopting Speech Recognition. *Commun. ACM*. 47(1), 69–75.

Didiot, E., Illina, I., Fohr, D. and Mella, O. (2010). A Wavelet-Based Parameterization for Speech/Music Discrimination. *Computer Speech and Language*. 24(2), 341–357.

Farooq, O. and Datta, S. (2004). Wavelet based robust sub-band features for phoneme recognition. *IEEE Proceedings Vision Image and Signal Processing*. 151(3), 187–193.

Flynn, R. and Jones, E. (2008). Combined Speech Enhancement and Auditory Modelling for Robust Distributed Speech Recognition. *Speech Communication*. 50(10), 797–809.

Flynn, R. and Jones, E. (2012a). Feature selection for reduced-bandwidth distributed speech recognition. *Speech Communication*. 54, 836–843.

Flynn, R. and Jones, E. (2012b). Reducing bandwidth for robust distributed speech recognition in conditions of packet loss. *Speech Communication*. 54(7), 881–892.

Furui, S. (1986). Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 34(1), 52–59.

Furui, S. (2010). *History and Development of Speech Recognition*, Springer US. 1–18.

Gaikwad, S. K., Gawali, B. W. and Yannawar, P. (2010). A Review on Speech Recognition Technique. *International Journal of Computer Applications*. 10(3), 16–24.

Gamulkiewicz, B. and Weeks, M. (2003). Wavelet Based Speech Recognition. In *IEEE 46th Midwest Symposium on Circuits and Systems, 2003*, vol. 2. 678–681.

Gandhi, T., Panigrahi, B. K. and Anand, S. (2011). A comparative Study of Wavelet Families for EEG Signal Classification. *Neurocomputing*. 74(17), 3051–3057.

Gilbert, M. and Feng, J. (2008). Speech and Language Processing Over the Web. *IEEE Signal Processing Magazine*. 25(3), 18–28.

Gowdy, J. N. and Tufekci, Z. (2000). Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. 1351–1354.

Graps, A. (1995). An introduction to wavelets. *IEEE Computational Science Engineering*. 2(2), 50–61.

Hecht-Nielsen, R. (1989). Theory of The Backpropagation Neural Network. In *International Joint Conference on Neural Networks (IJCNN)*, vol. 1. 593–605.

Jafari, A. and Almasganj, F. (2010). Using Laplacian eigenmaps latent variable model and manifold learning to improve speech recognition accuracy. *Speech Communication*. 52(9), 725–735.

Juang, B. H. and Rabiner, L. R. (2005). *Automatic Speech Recognition- A Brief History of the Technology Development*.

Karnjanadecha, M. and Zahorian, S. A. (2001). Signal Modeling for High-Performance Robust Isolated Word Recognition. *IEEE Transactions on Speech and Audio Processing*. 9(6), 647–654.

Kinnunen, T. and Li, H. (2010). An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. *Speech Communication*. 52(1), 12–40.

LeCun, Y., Bottou, L., Orr, G. B. and Mller, K.-R. (1998). *Efficient Backprop*, Springer. 9–50.

Lippmann, R., Martin, E. and Paul, D. (1987). Multi-style training for robust isolated-word speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 87).*, vol. 12. 705–708.

Loizou, P. C. and Spanias, A. S. (1996). High-Performance Alphabet Recognition. *IEEE Transactions on Speech and Audio Processing*. 4(6), 430–445.

Long, C. J. and Datta, S. (1996). WaveletBased Feature Extraction for Phoneme Recognition. In *Proceedings of the Fourth International Conference on Spoken Language, 1996. ICSLP 96.*, vol. 1. 264–267.

Mallat, S. G. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 11(7), 674–693.

McLoughlin, I. (2009). *Applied Speech and Audio Processing: With Matlab Examples*. (1st ed.). Cambridge University Press.

Negnevitsky, M. (2005). *Artificial Intelligence: A Guide to Intelligent Systems*. (2nd ed.). Harlow UK: Addison Wesley.

Nehe, N. and Holambe, R. (2012). DWT and LPC Based Feature Extraction Methods for Isolated Word Recognition. *EURASIP Journal on Audio, Speech, and Music Processing*. 2012(1), 7.

Nehe, N. S. and Holambe, R. S. (2008). New Feature Extraction Methods using DWT and LPC For Isolated Word Recognition. In *IEEE Region 10 Conference (TENCON)*. 1–6.

Nehe, N. S. and Holambe, R. S. (2009). New Feature Extraction Techniques for Marathi Digit Recognition. *International Journal of Recent Trends in Engineering (IJRTE)*. 2, 22–24.

Paliwal, K. (1992). Dimensionality Reduction of the Enhanced Feature Set for the HMM-Based Speech Recognizer. *Digital Signal Processing*. 2(3), 157–173.

Pandya, A. S. and Macy, R. (1996). *Pattern Recognition with Neural Networks in C++*. Florida: CRC Press.

Pavez, E. and Silva, J. F. (2012). Analysis and design of Wavelet-Packet Cepstral coefficients for automatic speech recognition. *Speech Communication*. 54(6), 814–835.

Peeling, S. M. and Moore, R. K. (1988). Isolated Digit Recognition Experiments using the Multi-Layer Perceptron. *Speech Communication*. 7(4), 403–409.

Picone, J. W. (1993). Signal Modeling Techniques in Speech Recognition. *Proceedings of the IEEE*. 81(9), 1215–1247.

Plannerer, B. (2005). *An Introduction to Speech Recognition*.

Razak, Z., Ibrahim, N. J., Idris, M. Y. I., Tamil, E. M., Yusoff, Z. M. and Rahman, N. N. A. (2008). Quranic Verse Recitation Recognition Module for Support in j-QAF Learning: A Review. *International Journal of Computer Science and Network Security (IJCSNS)*. 8, 207–216.

Rioul, O. and Vetterli, M. (1991). Wavelets and Signal Processing. *IEEE Signal Processing Magazine*. 8(4), 14–38.

Rosdi, F. (2008). *Isolated Malay Speech Recognition Using Hidden Markov Models*. Thesis.

Salam, M. S., Mohamad, D. and Salleh, S. H. (2009). Improved Statistical Speech Segmentation Using Connectionist Approach. *Journal of Computer Science*. 5(4), 275–282. 10.3844/jcssp.2009.275.282.

Salam, M. S. H., Mohamad, D. and Salleh, S. (2011). Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters. *The International Arab Journal of Information Technology*. 8, 364–371.

Sarikaya, R., Pellom, B. L. and Hansen, J. H. L. (1998). Wavelet Packet Transform Features With Application To Speaker Identification. In *Third IEEE Nordic Signal Processing Symposium*. 81–84.

Schafer, R. (2008). *Homomorphic Systems and Cepstrum Analysis of Speech*, Springer Berlin Heidelberg, chap. 9. 161–180. (Arden).

Shafik, A., Elhalafawy, S. M., Diab, S. M., Sallam, B. M. and El-samie, F. E. A. (2009). A Wavelet Based Approach for Speaker Identification from Degraded Speech. *International Journal of Communication Networks and Information Security*. 1(3), 52–58.

Shaughnessy, D. (2008). Invited Paper: Automatic Speech Recognition: History, Methods and Challenges. *Pattern Recognition*. 41(10), 2965–2979.

Vignolo, L. D., Milone, D. H. and Rufiner, H. L. (2012). Genetic wavelet packets for speech recognition. *Expert Systems with Applications*. (0).

Wu, J.-D. and Lin, B.-F. (2009). Speaker Identification using Discrete Wavelet Packet Transform Technique with Irregular Decomposition. *Expert Systems with Applications*. 36(2, Part 2), 3136–3143.

Xiao, X. (2009). *Robust Speech Features and Acoustic Models for Speech Recognition*. Ph.D. Thesis.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. (2000). *The HTK Book*. Microsoft Corporation.

Zhang, X.-y., Bai, J. and Liang, W.-z. (2006). The Speech Recognition System Based On Bark Wavelet MFCC. In *8th International Conference on Signal Processing*, vol. 1.