AUTOMATIC TRANSCRIPTION AND PHONETIC LABELLING OF DYSLEXIC CHILDREN’S READING IN BAHASA MELAYU

NIK NURHIDAYAT BINTI NIK HIM

SCHOOL OF COMPUTING
UUM COLLEGE OF ARTS AND SCIENCES
UNIVERSITI UTARA MALAYSIA
2015

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to :

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Pengecaman suara automatik *(ASR)* berpotensi untuk membantu kanak-kanak disleksia yang mengalami masalah pembelajaran. Kesalahan dalam penyebutan fonetik yang hampir sama oleh kanak-kanak disleksia amat tinggi sehingga memberi kesan kepada ketepatan pengecaman ASR. Oleh itu, objektif utama kajian ini adalah untuk menilai penerimaan ketepatan ASR dengan menggunakan transkripsi dan pelabelan fonetik automatik untuk kanak-kanak disleksia. Bagi mencapai matlamat utama tersebut, terdapat tiga objektif yang telah ditetapkan: pertama untuk menghasilkan transkripsi dan pelabelan fonetik manual; kedua untuk membina transkripsi dan pelabelan fonetik automatik menggunakan kaedah penjajaran paksa; dan ketiga untuk membandingkan ketepatan di antara transkripsi dan pelabelan fonetik automatik dengan transkripsi dan pelabelan fonetik manual. Lantaran itu, untuk mencapai matlamat kajian ini beberapa kaedah telah digunakan, termasuk pelabelan ucapan dan segmentasi manual, penjajaran paksa, *Hidden Markov Model (HMM)* dan Rangkaian Neural Buatan *(ANN)* untuk proses latihan, dan bagi mengukur ketepatan daripada ASR, Kadar Kesalahan Perkataan (*WER)* dan *False Alarm Rate* (FAR) digunakan. Sebanyak 585 fail ucapan telah digunakan untuk transkripsi manual, penjajaran paksa dan juga proses latihan. Pengecaman yang dijana oleh ASR enjin yang menggunakan transkripsi dan pelabelan fonetik automatik telah mencapai keputusan yang paling optimum iaitu 76.04% dengan kadar WER serendah 23.96% dan FAR 17.9%. Keputusan ini adalah hampir sama dengan ASR enjin yang menggunakan transkripsi dan pelabelan fonetik manual iaitu 76.26%, WER serendah 23.97% dan FAR 17.9%. Kesimpulannya, ketepatan daripada transkripsi dan pelabelan fonetik automatik adalah diterima bagi membantu kanak-kanak disleksia belajar menggunakan ASR dalam Bahasa Melayu (BM).

Kata Kunci: Pembacaan kanak-kanak disleksia, Transkripsi manual, Transkripsi dan pelabelan fonetik automatik, Penjajaran paksa, Pengukuran ketepatan ASR enjin.

Abstract

Automatic speech recognition (ASR) is potentially helpful for children who suffer from dyslexia. Highly phonetically similar errors of dyslexic children's reading affect the accuracy of ASR. Thus, this study aims to evaluate acceptable accuracy of ASR using automatic transcription and phonetic labelling of dyslexic children's reading in BM. For that, three objectives have been set: first to produce manual transcription and phonetic labelling; second to construct automatic transcription and phonetic labelling using forced alignment; and third to compare between accuracy using automatic transcription and phonetic labelling and manual transcription and phonetic labelling. Therefore, to accomplish these goals methods have been used including manual speech labelling and segmentation, forced alignment, Hidden Markov Model (HMM) and Artificial Neural Network (ANN) for training, and for measure accuracy of ASR, Word Error Rate (WER) and False Alarm Rate (FAR) were used. A number of 585 speech files are used for manual transcription, forced alignment and training experiment. The recognition ASR engine using automatic transcription and phonetic labelling obtained optimum results is 76.04% with WER as low as 23.96% and FAR is 17.9%. These results are almost similar with ASR engine using manual transcription namely 76.26%, WER as low as 23.97% and FAR a 17.9%. As conclusion, the accuracy of automatic transcription and phonetic labelling is acceptable to use it for help dyslexic children learning using ASR in Bahasa Melayu (BM).

Keywords: Dyslexic children's reading, Manual transcription, Automatic transcription and phonetic labelling, Forced alignment, Evaluation accuracy of ASR engine.

Acknowledgement

In The Name of ALLAH, Most Gracious, Most Merciful and Big Gratitude to Prophet, Muhammad S.A.W.

First and foremost, I thanked ALLAH the All Mighty for I am blessed to complete this study in time. Special thanks to Dr Husniza Binti Husni, my very helpful, supportive and dedicated supervisor for all her supervision, comments, ideas, suggestion and guideline given to me in order to complete this study.

My special thanks to Dr Mohd Hasbullah Bin Omar and Dr Norliza Katuk for the explanation and guidelines given to me especially during the preparation period and also during the presentation of this study. A special thanks also to all lecturers in Universiti Utara Malaysia for their great help and support during my academic career.

To my beloved family, a million thank you for their moral support and motivation especially my dad, Nik Him Bin Nik Ya and my mom Rohana Binti Kadir. Thanks for the love, encouragement, support and prayers. Last but not least, my fellow friends and others who have contributed directly and indirectly towards the completion of this study.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
ART	Automatic Reading Tutor
ASCII	American Standard Code for Information Interchange
ASR	Automatic Speech Recognition
BM	Bahasa Melayu
C	Consonant
CV	Consonant Vowel
CALL	Computer-assisted language learning
CoLiT	Colorado Literacy Tutor
CSLU	Center for Spoken Language Understanding
FAR	False Alarm Rate
HMM	Hidden Markov Model
HTK	Hidden Markov Modelling Toolkit
IPA	International Phonetic Alphabet
IRT	Interactive Reading Tutor
LD	Learning Disability
MDR	Miscue Detection Rate
MS	Milliseconds
NN	Neural Network
SER	Sentences Error Rate
TTS	Text to Speech
V	Vowel
WER	Word Error Rate

CHAPTER ONE
INTRODUCTION

1.1 Introduction

Automatic speech recognition (ASR) has been an essential technology, and it has come to a stage where it has been actively applied in a lot of industrial and consumer applications. ASR research is still in early stage in Malaysia for Bahasa Melayu (BM). However, ASR can play an important role in the education field like to boost children's is interest in learning. The availability of ASR technology gives opportunity to help children especially dyslexics to enhance their learning ability by using Automatic Reading Tutor (ART) or Interactive Reading Tutor (IRT). In order to develop ART and IRT using ASR technology, speech files of dyslexic children's reading aloud are used to perform transcription and phonetic labelling that serve as important basic elements for the construction of ASR engine (Athanaselis, Bakamidis, Dologlou, Argyriou, & Symvonis, 2014; Taileb, Al-Saggaf, Al-Ghamdi, Al-Zebaidi, & Al-Sahafi, 2013; Pedersen & Larsen, 2010; Husniza & Zulikha, 2009; Li, Deng, Ju, & Acero, 2008; Chuchiarini & Strik, 2003).

Since transcription and phonetic labelling are used for ASR engines, so the training and evaluation accuracy of it must be done by using standard methods and metrics (e.g. hybrids Hidden Markov Model (HMM) and Artificial Neural Network (ANN) for training; Word Error Rate (WER) and False Alarm Rate (FAR) for measuring accuracy). However, in this study the dyslexic children's speech presents a challenge to perform transcription and phonetic labelling due to dealing with highly phonetically similar errors that affected the performance accuracy of ASR engine. Thus, demonstrating the accuracy of transcriptions and phonetic labelling should be done.

The investigation of performance accuracy starts with producing transcription and phonetic labelling manually and automatically. Based on previous studies researchers believe that, the accuracy when using manual transcription and phonetic labelling is most accurate (Goldman, 2011; Dupuis, 2011; Yu, Gales, Wang, & Woodland, 2010; Dinarelli, Moschitti, & Riccardi, 2009; Hazen, 2006; Bauer, Hitzenberger, & Hennecke, 2002). This is because the procedure of manual transcription requires human transcribers to hear a sound of each phoneme before performing transcription and phonetic labelling make it more accurate compared automatic transcription and phonetic labelling. Even though, manual transcription has shown remarkable accuracy of spoken utterances, the accuracy performance of automatic transcription and phonetic labelling is still need to investigate. This is because the limitations in manual transcription and phonetic labelling to processing speech files are time consuming, costly and prone to error if involved thousands of speech files, researchers opted for transcription and phonetic labelling of speech through automated approach (Yuan, Ryant, Liberman, Stolcke, Mitra, & Wang, 2013; Husniza, Yuhanis, & Siti Sakira, 2013a; Schuppler, Ernestus, Scharenborg, & Boves, 2011; Van Bael, Boves, Heuvel, & Strik, 2007; Hosom, 2002).

The use of automatic transcription and phonetic labelling in transcribe and labelling of speech is now pervasive as the considerable gains in time and cost of automatic transcription made it an alternative way to handle limitation of manual transcription (Yuan et al., 2013; Cangemi, Cutugno, Ludusan, Seppi, & Van Compernolle, 2011; Kaur & Singh, 2010). Equally important, automatic transcription and phonetic labelling is the auto generated process to transcribe and label the speech signal into small units called phonetic symbols. This alternative approach can be performed faster compared to manual transcription (Silber & Geri, 2014; Cangemi et al., 2011; Sperber, 2012; Williams, Melamed, Alonso, Hollister, & Wilpon, 2011). Thus, automatic transcription and phonetic labelling approach are convenient to transcribe larger speech files and can avoid carelessness by the human transcribers like saving the files using the wrong filename or duplicating the files to name a few.

In this study, forced alignment is used to perform transcription and phonetic labelling and equally important, it is related with hybrid HMM/ANN as the state-of-the-art methods (Lu, Ghoshal, & Renals, 2013; Necibi & Bahi 2012; Novotney & Callison, 2010; Ting, Hussain, Tan, & Ariff, 2007; Hosom, 2002). This approach, adopted from ASR, is most widely used for transcription in speech synthesis, providing consistent and accurate speech segmentation (Ting et al., 2007). Then, the forced alignment task is applied Viterbi algorithm to find out the most accurate phonetic symbol for automatic transcription and phonetic labelling. Usually, before forced alignment process, a search for the suitable phonetic symbols and locations for all boundaries, the small amount of manual transcription and phonetic labelling must be train using hybrid HMM/ANN to ease Viterbi algorithm searching and learning featured of preferred symbol and boundary location.

In this work 585 speech files has been completed by automatic transcription and phonetic labelling using forced alignment and with similar amount of 585 speech files were perform using manual transcription. The hybrid HMM/ANN is essential to get accuracy of ASR together with construct ASR engine with 585 files transcription and phonetic labelling of manual and automatic transcription and phonetic labelling were be train. Last but not least, the accuracy performances of ASR engine are evaluated through WER and FAR. The acceptable accuracy of ASR engine using automatic transcription and phonetic labelling have been seen depend on result of manual transcription. Consequently, a comparison of automatic transcription and phonetic labelling against manual one need to be performed in order to see if it is at par with manual transcription.

1.2 Problem Statement

Dyslexia concerns with difficulty in reading, spelling and writing and thus regarded as learning disabilities (LD). Dyslexia happens due to disorder in the language processing parts of the brain. The pronunciation and reading properties of dyslexic children such as inability to interpret symbol correctly that affect their learning performance. A very significant problem that occurs on children with dyslexia is that they tend to produce highly phonetically similar errors (Husniza, 2010).

The most frequent error identified are highly phonetically similar errors such as when dyslexic children confusing sounds produced while they are reading aloud (Perea, Jimenez, Suarez, Fernandez, Vina, & Cuetos, 2014; Bourassa & Treiman, 2003). This is because the basic cause of dyslexia is a phonological weakness in processing the sounds of a language (Handler & Fierson, 2011). For example, the high phonetically similar errors usually occur when reading or spelling in BM for words such as *duku* instead of *buku*. The letter ‘b’ and ‘d’ are very similar in appearance and phonetically and thus create difficulty for dyslexic children to spell words as they may also reverse the order of two letters. Thus, dyslexic children’s reading is more difficult to recognize at phoneme level and it could be affected automatic transcription and phonetic labelling and possibly reduced accuracy of ASR while training.

Furthermore, in this work forced alignment is an approach to perform automatic transcription and phonetic labelling based on target words of lexical model. If the reading speech of dyslexic children’s contained highly phonetically similar errors, forced alignment also might have difficulty to transcribe and label phonetic symbols of the read speech, e.g. target word *maklumat* but dyslexic children’s read *malumat.* As a result, this can affect on accuracy of ASR engine using automatic transcription and phonetic labelling. Hence, reading failure of dyslexic children that contain highly phonetically similar errors also affected the accuracy of ART or IRT if they using this application. This is because of the nature of reading made by dyslexic children’s is difficult for these applications to recognize their speech due to confusing sounds speech produced during they are reading aloud. For example for word “*apa”* but the system recognized “*apah”* due to sounds of words are very similar in terms of articulation that makes ART or IRT system possibly to fail to recognize the words.

Thus, it is important for us to investigate the accuracy of automatic transcription and phonetic labelling of dyslexic children's read speech in BM using forced alignment to see whether its accuracy is acceptable that allows researchers to use automatic transcription and phonetic labelling for the purpose development ART and IRT for them.

1.3 Research Question

In investigating the accuracy of automatic transcription and phonetic labelling using forced alignment of dyslexic children's reading aloud in BM, can automatic transcription and phonetic labelling produce acceptable accuracy when dealing with highly phonetically similar errors of dyslexic children's reading?

1.4 Research Objectives

The main objective of this study is to evaluate acceptable ASR accuracy when using automatic transcription and phonetic labelling for dyslexic children's read speech as input files. To achieve this, the following sub-objectives have to be fulfilled:

i. To produce manual transcription and phonetic labelling.

ii. To construct automatic transcription and phonetic labelling using forced alignment.

iii. To compare between automatic transcription and phonetic labelling and manual transcription and phonetic labelling using WER and FAR as metrics.

1.5 The Scope

The scope of this study concerned with transcription and phonetic labelling of dyslexic children's speech. The transcription and phonetic labelling involved two approaches which are manual transcription and automatic transcription. The manual transcription and phonetic labelling task is accomplished using a tool called "Speech Viewer", which displays waveform, its spectrogram and any phonetic labels associated with it and allow them to be manipulated accordingly. Then, for automatic approach the transcription and phonetic labelling is done using forced alignment, prior to evaluation.

In this study the data collection using dyslexic children's read speech in BM obtained from previous research (Husniza, 2010). The 585 speech files data collection of 36 words was used for transcription and phonetic labelling that have been recorded from ten dyslexic children reading aloud in BM. Manual transcription acts as benchmark to investigate acceptance accuracy of automatic transcription and phonetic labelling using forced alignment. Thus, the performance of transcription and phonetic labelling starts through manual transcription process as transcribing speech file using manual approach is time consuming.

On the other hand, there are three important types of files necessary for accomplish automatic transcription and phonetic labelling task .wav, .txt and .phn. All these files are also needed for training process to develop ASR engines for both manual and automatic transcription and phonetic labelling for evaluation. The file format .wav is from the read speech recorded when dyslexic children's reading aloud single words in BM. Each speech file that has been recorded is saved in this format. Then, for format .txt file (text transcription of each word), it is created manually. For example, for a speech file of the word *cendawan*, a transcription file which contains its spelling saved as 'c-e-n-d-a-w-a-n' in DC-1.cendawan3.txt (filename). Then, for .phn format is produced after speech files completed transcribed and labelling speech files either using manual transcription or automatic transcription and phonetic labelling. The transcription and phonetic labelling is saved as *filename.phn* for phoneme file. In transcription and phonetic labelling, the phonetic symbols used are according to Worldbet (Hieronymus, 1993). Lastly, for the comparison of accuracy ASR, the standard metrics WER and FAR were used as they measured against target words and prior evaluation task, the transcription and phonetic labelling files were train using HMM/ANN.

1.6 Research Significant

In general, the focus of this study using automatic transcription and phonetic labelling by forced alignment would be valuable for ASR technology. This is because phonetic transcription and labelling that are done manually are known to face problems like subject to human error, tedious, costly and time-consuming. Instead, using automatic transcription and phonetic labelling, the task can be accomplished faster compared to using manual process to complete 585 speech file. Approximately three months are taken to finish manual transcription and phonetic labelling due to the problem magnified of dyslexic children’s read speech that dealing with highly phonetically similar errors.

Acceptable accuracy of dyslexic children’s speech using automatic transcription and phonetic labelling would be more advantageous to ASR technology. Moreover, this could save on cost to hire expert human transcribers to transcribe the thousands of speech files and label it phonetically. So, it can be concluded that automatic transcription and phonetic labelling is useful for many purposes, among which: faster transcription and phonetic labelling, gain time and cost, especially when large amounts of speech files are to be transcribed and labelled for bigger applications.

1.7 Research Overview

Table 1.1 presents the overall overview of the research. It outlines and maps the problem, research question, objectives, and methodology, as well as the expected deliverables.

Table 1.1. Research overview.

Problem statement	Research Question	Research objective	Methodology	Expected deliverables
Dyslexic children's speech contains highly phonetically similar errors that affect on its transcription accuracy.	Can automatic transcription and phonetic labelling produce acceptable accuracy when dealing with highly phonetically similar errors of dyslexic children's reading?	**Main-objective** To evaluate acceptable ASR accuracy when using automatic transcription and phonetic labelling for dyslexic children's read speech in BM.		
		Sub-objective a) To produce manual transcription and phonetic labelling.	Manual Transcription	Manual transcription and phonetic labelling of dyslexic children's reading (.phn).
		b) To construct automatic transcription and phonetic labelling using forced alignment.	Forced alignment	Automatic transcription and phonetic labelling of dyslexic children's reading (.phn).
		c) To compare between automatic transcription and phonetic labelling and manual transcription and phonetic labelling using WER and FAR.	i) ASR training method HMM/ANN ii) Alignment method WER & FAR	ASR accuracy of automatic transcription and phonetic labelling. ASR accuracy of manual transcription and phonetic labelling.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

ASR system allows us to communicate using words spoken through microphone or other voice input device. Generally, ASR system is defined as mechanism exchange speech from a recorded audio signal to a written equivalent of the message information e.g text (Shrawankar & Mahajan, 2013; Lakra, Prasad, Sharma, Atrey, & Sharma, 2012; Mandal, Das, Mitra, & Basu, 2011). According to Jurafsky and James (2000), a more technical definition defined that ASR as the building of system for mapping acoustic signals to a string of words. Normally, the ASR system implemented in the form of dictation software and intelligent assistants in computing, smart-phone or others device. There are two types of ASR systems which are speaker dependent ASR system that work by unit characteristics of a single person and depend on the speaker for training and speaker independent ASR system that's designed to recognize anyone’s voice, which no training is involved.

Today, ASR system is commonly used for many purposes like health care, military, telephony (e.g. smart-phones and customer helping) and education. In education field, ASR is important to help children with dyslexia to learning spelling and reading. Advances in ASR research have led to the development of ART or IRT for children’s education. However, ASR using BM language is still infancy and need for further research investigation to accomplish reasonable accuracy for certain tasks (Rosdi & Ainon, 2008; Ting, 2007).

In this study, the accuracy of ASR is used for statistically based system transforming speech signal dyslexic children reading aloud in BM into the corresponding sequence of linguistically defined units. The transcription and phonetic labelling speech database are useful for training, or at least initializing the static models for the selected units (Kvale, 1993). Therefore, this study focuses on transcription and phonetic labelling process because of transcription and phonetic labelling files are serving as basic element prior development ASR. However, highly phonetically similar errors of dyslexic children's read speech given challenges for its accuracy in this study.

Thus, this research review on challenges of dyslexic children's reading using BM language that present a Section 2.2. Then, overview of general architecture of ASR engine that have related with investigation transcription and phonetic labelling showed in Section 2.3. Hence, for Section 2.4 is discusses about issues using manual transcription and advantageous of automatic transcription and phonetic labelling performed their tasks. To perform transcription and phonetic labelling Section 2.5 provides suggestion methods to perform automatic transcription and phonetic labelling namely forced alignment, neural network and morphological phonetic transcription. Then, in section 2.6 prior evaluate accuracy of ASR using both manual and automatic transcription and phonetic labelling files were be train using hybrid HMM/ANN. Lastly, accuracy of ASR engines using both approaches were evaluated using dominant alignment metrics namely Word Error Rate (WER) and False Alarm Rate (FAR) in Section 2.7.

2.2 Challenges for Dyslexic Children Reading

Dyslexia is not a disease but it impedes children to learn to read, spell and write. The word dyslexia was introduced by Prof. Rudolf Berlin (Specialist and ophthalmologist) in 1887. Dyslexia came from a Greek word; 'dys' means difficulties and "lexia" means word (Newton & Thoman, 1974). He also has expanded the definition of dyslexia as the difficulty in language among children who are in process of learning and in ability to reading, writing or speaking with their other intellectual abilities (Newton & Thoman, 1974). Dyslexia is also having low reading skill. While reading or writing them often inventing the words and reverse the letters. Similarly landing the letters usually can't be differentiated by dyslexic children such as 'b-d", 'u-n', 'm-w', 'p-q' and 'b-p'.

According to Rello and Llisterri (2012) dyslexia is specific learning disabilities that roots from neurobiology. The dyslexic children's learning issues are classified into six characteristic (i) problem in concentrating, (ii) difficulty using the language, (iii) not be able write eloquently from board or book, (iv) the imbalance with intellectual ability, (v) tired eyes after concentrating on the writing for several minutes, and (vi) limited concentration (Lee, 2008).

DeFries, Olson, Pennington, & Smith (1991) attempted to explain the relationship between dyslexia with reading aspect. He said that there are four elements to speak such as experiences, listening, writing, and speaking. Dyslexic children who have involved in difficulty to read would impact several aspects contained in the readings.

For most normal human, reading seemingly used the vision, perception, recognition, understanding and reaction. Nonetheless, for children with dyslexia it was a challenging task to doing that. The main weaknesses of dyslexia lie in two aspects which are oral weakness and make phoneme discrimination. There are three essential aspects in discriminating phonemes like the sound aspect, phoneme aspect and speech aspect. The problem at the sounds stage is difficult to solve different auditory stimulus (linguistic and non-linguistic). This is caused by the frequent formation of sound in the dyslexic children's ears. The fundamental flaw in this stage was disrupting various phonological skills required for reading and spelling.

On the other hand, during reading process, dyslexic children couldn't recall letters, words, and even different sentences due to their eyes couldn't capture the reading material and causes to skip lines when read (Mohammad, Ruzanna, Vijayaletchumy, Aziz, Yasran, & Rahim, 2011). Because of this, children with dyslexia tend to produce highly phonetically errors (Carroll & Myers, 2010; Douklias, Masterson, & Hanley, 2010; Lee, 2008). For example, the harder letters to distinguish and often failure to pronounce that cause for highly phonetically similar errors like u/ and n, /m/ and /w/, as well /h/ and /l/. This situation create problem to these words 'masa'-'wasa', 'hari'-'lari', 'makan'-'makau'. The insertion of letter also happened to the word 'padang' when it is read as 'pandang'. Therefore, while children with dyslexia use ASR technology possibly their speech could affect the accuracy performance.

Previously, many researcher concentrate on the use of computer technologies to address the problem of dyslexia (Taileb et al., 2013; Athanaselis et al., 2012; Hagen, Pellom, & Cole, 2003; Russell, Brown, Skilling, Series, Wallace, Bonham et al., 1996). ASR technology can be used to support dyslexia learning in their daily task that potentially good for assisting children with dyslexia by training them to grasp literacy skill (Russell, Brown, Skilling, Series, Wallace, Bonham, & Barker, 2007). According to Conn and McTear (2000) suggested using ASR to help dyslexic children as well adults in writing process through dictation.

Taileb et al. (2013) develop an Arabic assistance solution for dyslexic children, it is ASR software based on analyzing phonetic isolated Arabic alphabet letters. This software application provides an environment for dyslexic children to develop and enhance their skill of reading and spelling. Besides, Athanaselis et al. (2012) present their work with effort to incorporate a state of the art speech recognition engine into new platform for assistive reading for improving reading ability of Greek dyslexic students. They reported that the platform was developed in the framework of the Agent-DYSL, IST project, and facilities dyslexic children in learning to read fluently. The performances of ASR technology usually similar because the basic component of ASR engine architecture is to recognized speech signal from user.

2.3 Overview of ASR Engine Architecture

ASR is the process by which used a machine (e.g. computer) has allowed user to recognize and act upon spoken language or utterances. Generally, ASR is performed by computer algorithm designed to take speech signal (waveform) as input and produce as output through certain processes. The processes of ASR engine rely on three core components, i.e. acoustic model, language model and pronunciation dictionary or know as lexicon model. Figure 2.1 shows the general of ASR system architecture that related to this study. The subsections forwards were describes the details of ASR system components.

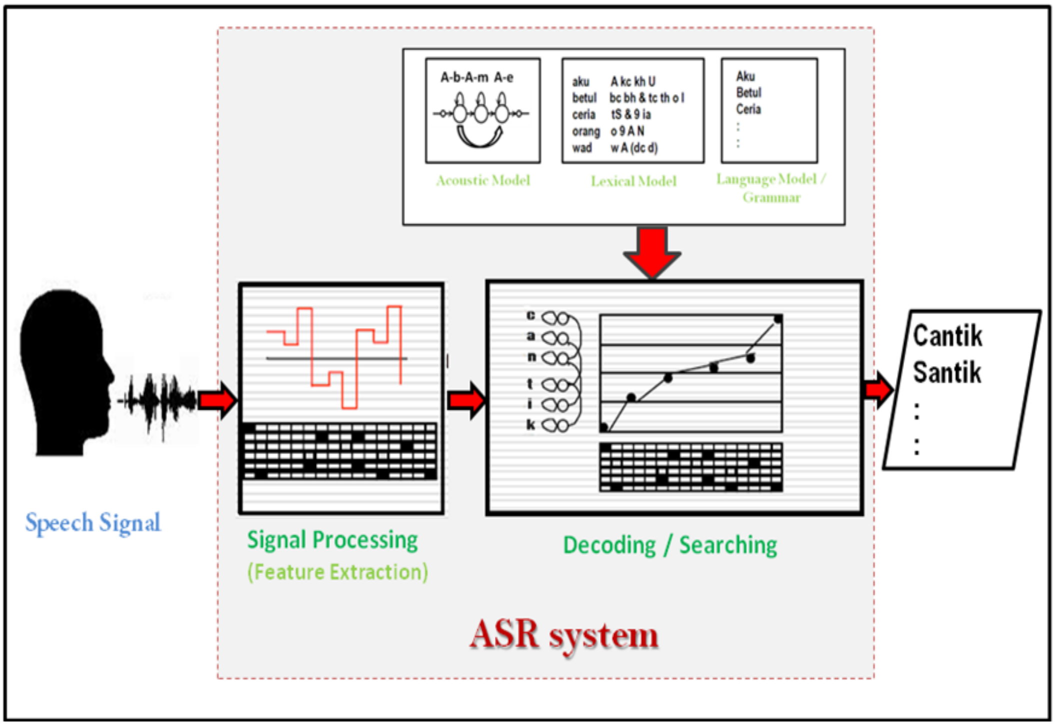


Figure 2.1. General of ASR system architecture.

2.3.1 Speech Signal

The ASR system was used in two recognition mode by using recognition of spoken words (word output) or recognition speech sounds or phone (phone output) (Zekveld, Kramer, Kessens, Vlaming, & Houtgast, 2008). Figure 2.1 shown that the processing chain starts with speech signal (left side), which for ASR system, speech signal can be served as input for training and testing. Speech signal can be present any word or languages in this world. Hence, in this work, the speech signal is from dyslexic children's reading aloud the single word in BM.

2.3.2 Signal processing

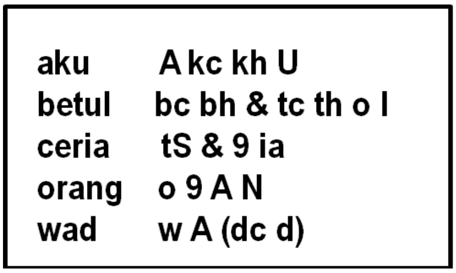
Speech signal processing is the initial stage of ASR recognition, it is through this approach that the system views the speech signal itself. Theoretically, feature extraction possibly to recognize speech directly from the digitized waveform (speech signal). Feature extraction incorporate knowledge of the nature of speech sounds in measurement of the features and utilize rudimentary models of human perception (Picone, Ganapathiraju, & Harmaker, 2007).

2.3.3 Acoustic Model

An acoustic model is created by taking audio recordings of speech, then their text transcriptions is using software to create statistical representations of the sounds that make up each word (Yang, Oehlke, & Meinel, 2011). It is used by an ASR engine to recognize speech. The ASR engine process required acoustic model to recognized the speech sounds because it can create transcription which taken from speech corpus and compiling them into a statistical representations of the sounds. This activity is through a training process using hybrid HMM/ANN. Similarly to other language, acoustic model of BM means pronunciation modelling i.e., mapping of the lexical units to phone like units or acoustic modelling of the basic phone units.

2.3.4 Lexical Model

BM is the official language of the Malaysian, Indonesia and Brunei. This language is part of Austronesia language that used for education system in Malaysia. There are some similar features between the BM language and English language. Commonly, Malay language is the structure well defined and can be unambiguously derived from a string. The basic syllable structure of the BM is generated by ordered of three syllabication rules which are Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) are the most familiar and easily to be found almost in every Malay primary words. Sub-words unit requires for lexicon to describe the constituents of a word. The lexicon model is contains the vocabulary of words and theirs associated phone strings. For example lexicon words refer Figure 2.2 below:



aku	A kc kh U
betul	bc bh & tc th o l
ceria	tS & 9 ia
orang	o 9 A N
wad	w A (dc d)

Figure 2.2. Lexicon model in BM.

2.3.5 Language Model or Grammar

A language model has an important role in ASR system either perform transcription and phonetic labelling or training ASR engine. The language model or grammar file’s ensures that the translated words with valid linguistic sentence. Hence, language files model containing the probabilities sequence words and grammar is a smaller files that containing sets of predefined combination of words. Language model are used to calculate the a priori probability *P (W)* for given words sequences $W= w_1,w_2,w_3,w_4...w_n$. The probability much depends on of the acoustic signal X.

2.4 Transcription and Phonetic Labelling Performances

The transcription and phonetic labelling (or phonetic segmentation) determines the position of phoneme, all words boundaries and syllable in speech corpus of any duration on the basic of the speech files recoding and its orthographic transcriptions (Goldman, 2011). Aligned speech audio are widely used in many speech application including ASR, speech synthesis and phonetic research. Because of this, the transcription and phonetic labelling need to be emphasized prior to development of ASR engine. The transcription can be performed using two approaches either using manual transcription or automatic transcription and phonetic labelling. Then, the speech signal (audio recording) is used as input for both approaches. This work focuses to present the automatic transcription and phonetic labelling instead of manual transcription technique. However, in this study, manual transcription acts as benchmark to compared accuracy of automatic transcription and phonetic labelling because manual transcription is probably the possible approximation of the most accurate technique to transcribing speech files (Cangemi et al., 2011).

2.4.1 Manual Phonetic Transcription and Its Limitation

Manual transcription is the accurate and more reliable method of creating phonetic symbols and time-aligned labels by expert human transcribers (Husniza, Yuhanis, &

Siti Sakira, 2013b; Goldman, 2011; Kim & Gibbo, 2011; Mporas, Ganchev, & Fakotakis, 2010; Giurgiu & Kabir, 2012). However, the current practices is that researchers tend to use arbitrary human transcribers for transcribing speech files due to human expert transcribers is costly (Novotney & Callison-Burch, 2010; Hofmann & Pfister, 2010; Chuchiarini & Strik, 2003; Demuynck & Laureys, 2002). As William et al. (2011) states that historically transcription and phonetic labelling has been an expensive and slow process done by human transcribers and usually require at least 6 hours of work per hour of speech. Hence, in U.S.A the resulting cost for manual transcription is $90-$150 per hour of speech (Novotney & Callison-Burch; 2010; Milde, 2014; Passy, 2008; Kimball, Kao, Arvizo, Makhoul, & Iyer, 2004). Thus, the limitation manual transcription are costly, time consuming and convenient for transcribe and labelling in small speech files (Vasilescu, Vieru, & Lamel, 2014; Goldman, 2011; Dupuis, 2011; Yu, Gales, Wang, & Woodland, 2010; Dinarelli, Moschitti, & Riccardi, 2009; Hazen, 2006; Bauer, Hitzenberger, & Hennecke, 2002). Moreover, the problem is magnified if thousands of speech files need to be transcribed by human that tend to them became tedious and laborious. Thus, the redundancy of the speech filenames, human error during labelling and wrong transcription could potentially occur during transcription process.

Normally, generating transcription and phonetic labelling that done manually by using software tool like CSLU toolkit by Sutton, Cole, Villiers, Schalkwyk, Vermeulen, Macon, Cohen et al., (1998), Praat toolkit Boersma and Weenink (2013) or Wavesurfer toolkit Sjolander and Beskow (2006) that can displays spectrogram, speech waveform, phonetic labels, and other acoustic information. The CSLU toolkit of manual transcription approach is used in this study to perform manual transcription and phonetic labelling by listening to the sound of speech files (waveform) and at the same time using knowledge of the relationship between the waveform, i.e. spectrogram and phonetic contents to transcribe and align the word according to sound of phoneme. Ideally, the phonetic transcription is the use of phonetic symbols to represent the speech sound by written phonetic symbols according to phoneme (Wells, 2006). The manual transcription transforms a speech sound into distinct small units shown in Figure 2.3. One of the phonetic symbols inventories during transcription task is Worldbet phonetic symbol (Hieronymus, 1993). This Worldbet phonetic symbol can cover all of the world’s languages in a systematic manner.

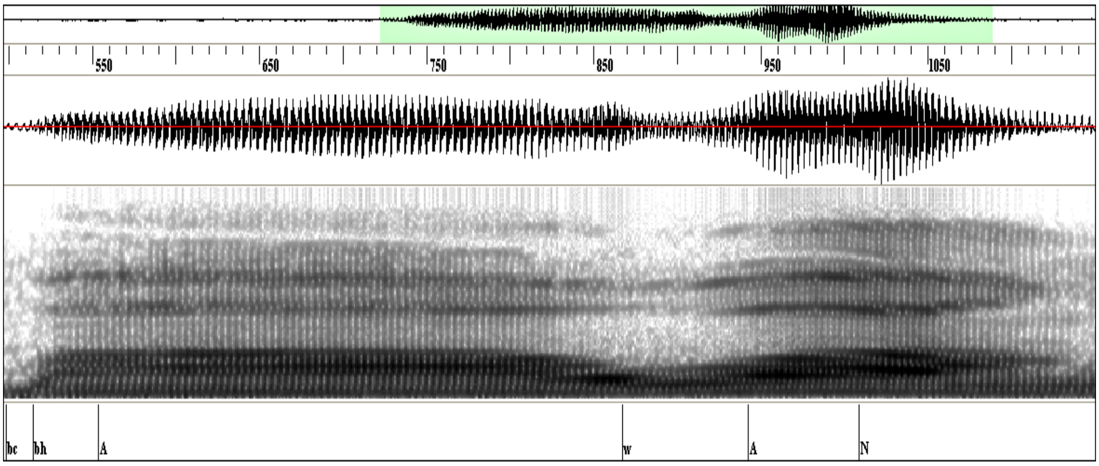


Figure 2.3. An example of manual segmentation and phonetic labelling for the word *“bawang”*.

Besides the aforementioned problem, the other issue during transcribing and labelling phonetic symbols that are done manually is complexity and laborious due to highly phonetically similar errors of dyslexic children’s reading (Husniza et al., 2013a; Chuchiarini & Strik, 2003). This is because, while listening to word to perform transcription and phonetic labelling of dyslexic children’s reading speech files, human transcribers often heard similar sounds. For example, listening to the words *ayat* and ‘ayah’ or *selesa* and ‘selasa’ or *wad* and ‘wap’. Thus, it might be confused to human transcribers to perform transcription and phonetic labelling. Therefore, automatic transcription and phonetic labelling is potential approach as alternative to solve this problem.

2.4.2 Automatic Transcription and Phonetic Labelling

The automatic transcription and phonetic labelling is the alternative option to overcome the issues with manual transcription because of it effective approach (Goldman & Schwab, 2014; Yuan et al., 2013; Husniza et al., 2013b; Brognaux, Roekhaut, Drugman, & Beaufort, 2012; Hofmann & Pfister, 2010; Das, Izak, Yuan, & Liberman, 2010; Al-Manie, Alkanhal, & Al-Ghamdi, 2009; Tolba, Nazmy, Abdelhamid, & Gadallah, 2005; Kawai & Toda, 2004). The important factor for having an automatic transcription and phonetic labelling is because human labelling and segmentation often subject to error due to fatigue with content of various styles of pronunciations. Therefore, the human transcribers tends to saving the files using wrong filename and duplicating the filename. In addition, this problem magnified when human transcribers faces the highly phonetically similar errors, it is not always guaranteed to produce exact and accurate transcription. Thus, an automatic transcription and phonetic labelling approach is highly desirable.

Some studies already reported that the benefit of using automatic transcription and phonetic labelling system is to improve ASR speech synthesis system (Riley, Byrne, Finke, Khudanpur, Ljolje, McDonough et al., 1999; Wester, 2003; Yang & Marten, 2000; Saraclar & Khundapur, 2004; Tjalve & Huckvale, 2005; Cucchiani & Strik, 2003). The advantage of using automatic transcription and phonetic labelling is when it comes to exploring large speech files (Van Bael et al., 2007). This is because the automatic transcription and phonetic labelling can reproduce of speech files which can performs transcription and phonetic labelling files quickly and effectively (Goldman & Schwab, 2014).

Nevertheless, before automatic transcription and phonetic labelling is being used for dyslexic children’s read speech, it is essential to know how accurate they are especially when dealing with highly phonetically similar errors speech. The challenge in this study is acceptable accuracy of automatic transcription and phonetic labelling. The automatic transcription and phonetic labelling is not accurate enough compared to human transcribers (Hosom, 2009; Goldman, 2011, Chou, Tseng, & Lee, 2002). Thus, for measuring acceptable accuracy performances of automatic transcription and phonetic labelling, the manual transcription is the reference is considered as an accurate approach in phonetic transcription and phonetic labelling. There are several approach in the following subsection are discovered for automatically transcription.

2.5 Methods for Automatic Transcription and Phonetic Labelling

Automatic transcription and phonetic labelling approach were used to label and segment phonetic constituents of speech sounds. Hence, the most of current automatic transcription and phonetic labelling are derived from several approaches for develop ASR. The automatic transcription task can be performed by applying forced alignment, neural network or morphological phonetic transcription approach.

2.5.1 Forced Alignment

The most common approach for automatic transcription and phonetic labelling is forced alignment methods (Stolcke, Ryant, Mitra, Yuan, Wang, & Liberman, 2014; Yuan et al., 2013; Leither, 2008). The previous researchers that has been used forced alignment to performs automatic transcription and phonetic labelling which are Sarma, Saharia, and Sharma, (2014); Husniza et al. (2013b); Jakovljevic et al. (2012); Hofmann & Pfister, (2010); Hosom, (2009); Sjolander, (2003); Rapp, (1995). Forced alignment is the process by which phonetic boundaries can be determined; usually this process uses automatic speech-recognition technique determines the location of phonemes in speech waveform, given only sequence symbols that represents the phonetic content of the speech (Yuan et al., 2013).

The forced alignment method is done by employing Viterbi algorithm and it has been used extensively in speech research for different topic, ranging from speech recognition to speech synthesis and phonetic analysis (Schuppler, Ernestus, Scharenborg, & Boves, 2011; Yuan et al., 2013; Jakovljevic, Miskovic, Pekar, Secujski, & Delic, 2012; Hofmann & Pfister, 2010). Subsequently, the forced alignment enable to align the transcribed speech data with identifying which time segments in the speech data correspond to particular words or phoneme in the transcription data. The Viterbi algorithm is about finding a dynamic programming algorithm for searching a sequence of hidden states that called Viterbi path that results in a sequence of observed events as indicated by Fang (2009). The Viterbi algorithm finds the state sequences through the model and almost produced the sequence of feature vectors under consideration.

Forced alignment has the ability to recognize variation or lexicon model comprises utterance variant of word. The Viterbi algorithm can find the best matching of the given various pronunciations. This is supported by Jiang, Yuan, Tsaftaris, and Katsaggelos (2011); Lee, Katsamanis, Black, Baucom, Georgiou, and Narayanan (2011); Naghibi, Hofman and Pfister (2013); Cucchiani and Strik (2003), Demuynck and Laureys (2002), where the performance of the automatic phonetic transcription using this approach appears to be similar to that of the human transcribers. Hence, forced alignment also enable to facilitate speech recording (transcription and phonetic labelling) by providing an efficient and faster approach to segment and label the speech files prior to developing an ASR.

The forced alignment method is employed to train the speech files and extract speech features to produce transcription and phonetic labelling files. The main goal of using forced alignment is to represent the orthography of the spoken words (Kuo, Li &

Wang, 2007; Changemi et al., 2011; Martens, Binnenpoorte, Demuynck, Van Parys, Laureys, Goedertier, et al., 2002). In addition, forced alignment is an effective approach that can automatically generate phonetic transcription for a large number of speech files rather than using manual transcription that is normally for convenient small number of speech files (Hofmann & Pfister, 2013; Goldman, 2011).

Yuan and Liberman (2011) investigate the use of forced alignment for automatic transcription of "g-dropping" in American English. The g-dropping refer to phenomenon English where two acoustic models were trained, one for in' and the other for ing'. For example g-dropping that use of an apostrophe in place of g, such as in word nothin' and walkin'. The researcher selected randomly 200 words from Buckeye Corpus. The model for in' and ing were added to the Penn Phonetic Lab Forced Aligner and then forced alignment will choose probable pronunciation. The experiment "g-dropping" participant by ten Mandarin Chinese speakers and eight native American speaker using the same words. The agreement rates between the forced alignment methods for native English speakers ranged from 79% to 90%.

Phonetic transcription is an essential research issue in speech processing as represent of a speech sounds. Sarma, Sahariah, and Sharma (2014) transcribed Assamese Speech corpus using forced alignment. Assamese is an Eastern Indo-Aryan language and a less computationally aware language. The speech files were transcribed of two type symbols which are 38 phonemes for ASCII transcription and 34 phonemes for IPA transcription. For Assamese speech it is difficult to pronounce due to similar words is pronounced with different sound of language. The researcher reported that till now it has not been possible to get 100% accuracy for any languages and the accuracy result from their experiment is only 65.26%.

Vijayalaksmi (2012) study has used Malay language to perform transcription and phonetic labelling using force alignment where she implemented force alignment algorithm that made itself. She reported that Malay language as known as BM and the English language has several similarities in term of phonetic language. Then, both Malay and English wrote using the roman characteristic. In their project, her implemented force alignment for transcribing and label 3437 words in BM that consisted 44 phonemes. However, the data her use using normal speaker which not dealing with highly phonetically similar errors. Thus, the result of the automatically segmentation using force alignment is 95.7% with boundary ranges differ less than 30ms.

Subsequently, the previous work by Hosom (2002) makes comparison between manual transcription and automatic transcription to determine the quality accuracy of ASR recognizer. This experiment using OGI 30K Numbers corpus was collected from thousands of people (e.g. telephone number, street address, zip code and etc.) over the telephone natural speaking. In order to compare the proposed techniques, the automatic alignments were obtained from best general-purpose forced alignment system and the manual alignments generated by expert labellers. These both recognizers of transcription and phonetic labelling were trained and evaluate using the same dataset of telephone-channel continuous speech. Manual alignment is reported to have 97.54% for word accuracy and 90.18% sentences accuracy. Then, the result of automatic alignment is similar with manual which is 97.24% word accuracy and 88.8% sentence accuracy. These results percentages of accuracy two recognizers are higher because of they are not involved highly phonetically similar errors of dyslexic children's reading.

Besides this previous work, there are several researchers has been done using forced alignment approach to performs automatic transcription and phonetic labelling. Table 2.1 shows performances automatic transcription and phonetic labelling of different studies and different languages.

Table 2.1. Performances automatic transcription and phonetic labelling using forced alignment of different studies.

Sources	**Corpus**	**Performances**
Yuan & Liberman 2011	English Corpus	79.0%-90.0%
Cangemi et al., 2011	Italian	94.0%-20 ms 97.0%-30 ms
Yuan et al., 2013	TIMIT (54 Phonemes)	93.92%
Kuo & Wang, 2006	TIMIT acoustic phonetic continuous speech corpus	71.1%
Lin, Jang, & Chen, 2005	Mandarin Chinese Corpus	72.1%- 20 ms 87.4%- 30 ms
Sjonlander, 2003	Swedish	85.5%

Note that these works present performances of normal speech without consideration of any errors. However, their performances indicated that forced alignment (fa) can be used for automatic transcription and phonetic labelling of dyslexic children read speech.

2.5.2 Neural Network

Neural network is another approach to perform automatic transcription and phonetic labelling. This approach allowed training recognizer with the forward and backward algorithm. Although neural network is not the dominant technique in automatic transcription and phonetic labelling, Chang, Shastri, and Greenberg (2000) has developed it to label and segment phonetic constituents of spontaneous American English using neural network. The system includes a Viterbi-like decoder and two neural network stages. The first neural network performed classification of each frame along five articulatory-base dimensions (the place of articulation, manner of articulation, voicing, lip-rounding, front-back articulation). These phonetic features go through to the second neural network that maps onto phonetic-segment labels. The preliminary results from this performance Neural Network using American English corpus database within 10 ms is 82.5%.

The Neural Network approach also known as Multi-layer perceptron was used by Togneri, Alder, and Attikiouzel, (1990). A three layer perceptron network was used to classify the /i/ sound of isolated words from different speaker. The neural network has been used to identify different phoneme. The phonotopic map provides techniques for segmenting speech sounds into their basic units. The result the classification of speech that has been achieved of Togneri et al. (1990) is 97%.

Kawachale and Chitode (2012) present methods for automatic speech signal segmentation using neural network. They state that automatic transcription and phonetic labelling is required because manual transcription is extremely time consuming and also some restrictions of human transcribers limitation. In conducted transcription and phonetic labelling using neural network MAXNET and k-means algorithm are adopted. MAXNET is one layer neural network that present competition to determine which node has the highest input value. The speech in Indian language is based on basic sounds units which are inherently syllable unit from consonant (C), Vowel (V), and consonant vowel consonant (CVC) combination. Based on her study, a database of around 4000 to 5000 words has been prepared for the transcription and labelling using neural networks. Around 90% accuracy is achieved with neural network models for syllable transcription which resulted in naturalness improvement of Marathi Text to Speech (TTS) (Kawachale & Chitode, 2012).

2.5.3 Morphological Phonetic Transcription

The transcription and phonetic labelling of speech files also can be done using morphological phonetic transcription that can perform automatically. The system present by Wothke (1993) created transcription and phonetic labelling from the orthographic representation in order to provide multiple pronunciations to training the speech recognizer and to generate phonetic transcriptions it does not required speech signal.

Previous work adopts morphological phonetic transcription to transcribe German as the target language (Leither, 2008). In German the pronunciation of a letter depends on the morphological context thus the simple application of rules tend to be erroneous. Hence, the proposed solution for this limitation is a transcription into morphology i.e suffix, stem and prefix. The transcription and phonetic labelling was created after the first step a letter-to-phone mapping is applied. The morphological phonetic transcription performed for each transcription of word with identical orthographic representations but different morphology is created. So, the possible transcription and phonetic labelling for multiple pronunciations are considered and much depends on phonetic transcription are generated. The tested result using morphology phonetic transcription by Leither, (2008) is 92.2% of the words in a test set the system and this approach used to cover a part of the process of automatic transcription of an actual utterance. Hence, this method provided simple rule-base transcription that led to enhanced transcription generation.

2.6 Training ASR Engine

In training, typically transcription and phonetic labelling files were used to develop ASR engine together with produce accuracy of ASR. The accuracy of ASR is essential because of it is use to measure accuracy performances of ASR. In order to training for develop ASR engine and present accuracy of ASR, many methods can be used such as hybrid HMM/ANN, ANN, Dynamic Time Warping (DTW), Semi-continuous HMM (SCHMM) and Vector Quantization (VQ). However, the most dominant methods used for training ASR normally researchers applying hybrid HMM/ANN (Frikha & Hamida, 2012; Ong & Ahmad, 2011; Gemello, Mana, & Albesano, 2010). Furthermore, in this work hybrid HMM/ANN are used due to it comes from complete packages to present training ASR engine of CSLU toolkit software.

The hybrid HMM/ANN suggests building the lexical model and language model to emphasized towards recognition accuracy. With the training of transcription and labelling files using hybrid HMM/ANN, ASR technology can be growing rapidly. In addition, ANN is a fault tolerance and nonlinear property when applied in ASR using neuron network (Haykin, 1999; Azam et al., 2007). Numerous research and development of ASR engine have been done in recent years using HMM and ANN as the hybrid method outperformed from other methods (Frikha & Hamida, 2012; Ong & Ahmad, 2011; Husniza & Zulikha, 2009; Fadhilah & Ainon, 2008; Bourland & Morgan, 1994). Ong and Ahmad (2011) proposed to use hybrid HMM/ANN method for developing a speaker independent and Malay speech recognition. Table 2.2 depicted performances using hybrid HMM/ANN, Vector Quantization (VQ), Support Vector Machines (SVM) and other methods in different corpus.

Table 2.2. Review result accuracy of different speech recognizer.

Source	**Speech Recognizer**	**Corpus**	**Performance**
Ong & Ahmad 2011	Semi-continuous Hidden Markov Model (SCHMM)	BM, adult speech	99. 66 %
	Hidden Markov Model (HMM) / Artificial Neural Network (ANN)	BM, adult speech	100 %
Fadhilah & Ainon 2008	Mel Frequency Cepstral Coefficients (MFCC) 5 states HMM with mixture Gaussian densities Baum Welch algorithm for training	Isolated 6 Malay vowels Adult “Empat”, “Lapan”, “Rekod‟, “Tidak”, “Tujuh” & “Tutup”	88.67 %
Ting et al., 2007	Artificial Neural Network (ANN)	Isolated word in BM	92.00 %
Marcus et al., 1993	Hybrid HMM/ ANN	English corpus	96.33 %
Mohamad et al., 2011	LPC Error BP used to improve NN	Isolated words Malay digits 0-9	97.67%

Abushariah, Gunawan & Khalifa, 2010	Hybrid HMM/ANN	English digits from (Zero through Nine)	99.50%
Jackson, 2005	Hybrid HMM/ANN	Kinyarwanda Language	94.47%
Choudhary, Chauhan & Gupta, 2010	Hidden Markov Modeling Toolkit (HTK)	Hindi language	90.00%
Bhotto & Amin, 2004	Vector Quantization (VQ)	“Bangali Text Dependent Speaker”	70.00% to 85.00%

Note that these research present performances of training using normal speaker of different languages. Therefore, their performances obtained high accuracy.

The most researchers are familiar with hybrid HMM/ANN for isolated word task with sufficient training data, each word is trained by single hybrid HMM/ANN (Toth & Kocsor, 2007; Ramesh & Gahankari, 2013). So, in this study, training ASR engine for dyslexic children’s reading were flexible and convenient using this method. Furthermore, the hybrid HMM/ANN is the preferable method to use due it’s simpler, faster, and flexible for modelling and training larger speech files (Rabiner & Juang, 1993). After training the automatic transcription and phonetic labelling of dyslexic children’s speech with hybrid HMM/ANN, the ASR was evaluated in terms of its accuracy using WER and FAR as metrics.

2.7 Evaluation of ASR Accuracy

To evaluate any ASR for its accuracy, the ASR has to be trained first. The training can be performed using hybrid HMM/ANN as the state-of-the-art methods. Accuracy of ASR required determining the quality of transcription and phonetic labelling. Currently, this is done by comparing the automatic transcription and phonetic labelling with manual transcription as a reference for measuring acceptance accuracy (Yuan et al., 2013; Radi 2013; Hosom, Shriberg, & Green, 2004; Barras, Geoffrois, Wu, & Liberman, 2002; Saraclar & Khundanpur, 2004). Before automatic transcription and phonetic labelling can be used for ASR, it is important to know how accurate they are. The concept of accuracy is expressed as a percentage of the number of words that is recognized correctly out of total number of words spoken (in this case, the read words) (Pieraccni, 2012).

The accuracy of ASR technology is uses to recognize dyslexic children reading selected vocabulary of isolated word in BM. There are several of metrics to evaluated accuracy of transcription and phonetic labelling which are using WER and FAR. Then, besides WER and FAR the alignment metrics for evaluation accuracy of ASR are Miscue Detection Rate (MDR), Sentence Error Rate (SER), Digit Error Rate (DER) and Semantic Error.

2.7.1 Word Error Rate

Word error rate (WER), is the most dominant metric of the performance of speech recognition or machine translation system (Mishra, Ljolje, & Gilbert, 2011; Yoon, Chen, & Zechner, 2010; Husniza, 2010; Fish, Hu, & Boykin, 2006; Wang, Acero, & Chelba, 2003). Generally, WER was used to measure ASR accuracy and penalizes all types of ASR error. According to Hagen (2004), the WER is the highly valid metric that is widely accepted and easy to use. So, it can be a useful measurement accuracy of ASR. The quality of the output transcription and phonetic labelling of ASR typically depends on WER metric where the lower percentage the better accuracy. WER can be computed as:

$$\boxed{\textbf{WER} = \ \ \textbf{100} - \textbf{Recognition Rate \%}}$$

The WER threshold for acceptable performance is different for dissimilar application. It has been shown a good performance for previous study in evaluate quality of the output transcription and phonetic labelling of ASR. The examples of studies that use WER to measure and evaluate their ASR performance are by Lee, Hagen, Romanyshyn, Martin, and Pellom (2004) who managed approximately 30% of the word error, Shire (2001) with a low WER of 7.3% and Hagen et al. (2003) with average WER of 27.87%. Rahman, Mohamed, Mustafa, and Salim (2014) in their quest to explore the ASR technology for Malay speaking children have found promising WER score of 23.30% using children speech at the word level.

Since this study concerns with dyslexic children’s reading with highly phonetically similar errors, the optimum rate for WER is defined to follow that of state-of-the art phoneme recognition rate between 70% to 75% as suggested by Hosom (2007). This means that the optimum rate for WER is defined to range from 25% to 30%. The metric used together with WER is False Alarm Rate (FAR) because is not enough for us only using WER to evaluating accuracy transcription performances. Lee et al. (2004) support this notion as WER alone do not provide any diagnostic information. For this reason, FAR are also used in this study along with WER to evaluate performances of the ASR engine using manual transcription and ASR engine using automatic and phonetic labelling.

2.7.2 False Alarm Rate

The evaluation accuracy of ASR engine also employs False Alarm Rate (FAR). The FAR Rate is “erroneous radar target detection decisions caused by noisy environment or other interfering signals exceeding the detection threshold”. The challenge task in this research is whether or not the ASR is accurate enough when trained using automatic transcription and phonetic labelling that produced from dyslexic children’s reading. According to Mostow (2006), FAR is the “percentage of correctly read words rejected by ASR”. Therefore, FAR provides richer information as how accurate an ASR system recognizing correct read words.

2.7.3 Miscue Detection Rate

The field of miscue detection rate (MDR) and confidence scoring is large and a lot of research have been conducted within the area to made evident survey (Jiang, 2005; Banerjee, Beck, & Mostow, 2003; Rasmussen, Tan, Lindberg, & Jensen (2009). Dyslexic children's speech is differing significantly from normal children read speech in a number of ways. According to Pedersen and Larsen (2010) some of the MDR encountered in dyslexic children' read speech such as regression, filled pauses, word skipping, word truncation and long pauses between words. Thus, the MDR metric is suitable to define as the number of miscues which have not been detected divided by the number of all miscues. The real information (performances) MDR the high accuracy percentage is better for achievement recognition (Banerjee et al., 2003; Li, Ju, Deng, & Acero, 2007).

2.7.4 Sentence Error Rate

The sentence error rate (SER) is similar with WER but it indicates the percentage of sentences, not percentage of words which is calculated by comparing the hypothesis string generated by the decoder to the reference string and scoring the whole sentence as wrong if they differ (Evermann, 1999). SER shows similar advantages and shortcomings as WER. More about SER is the edit distance between references word sequences. The edit distances can defined as the minimum number (or weighted sum) of substitution (Sub), deletions (Del), and insertions (Ins) to transform one string to an others. Thus, to comprehend about Sub, Del and Ins see example below that illustrated the comparison actual sentences as a references and error sentences as a hypothesis.

i) Substitution

Ref: You are wrong.

Hyp: You are strong.

A substitution happened. “Wrong” was substituted by “strong” by the ASR.

ii) Insertion

Ref: Where is the river?

Hyp: Where is the Amazon river?

An insertion happened. “Amazon” was inserted by the ASR.

iii) Deletion

Ref: I love pink shoes.

Hyp: I love pink.

A deletion happed. “Shoes” was deleted by the ASR.

Then, SER can be computed accuracy of ASR in sentences level as:

$$\boxed{\textbf{SER} = \quad \textbf{Edit Distance * 100}}$$

Where Edit Distance: (S+I+D) / N

S is the number of incorrect words substations, I is the number extra words insertion, D is the number deletion and N is the number of word that correct recognize. While we adopt the dominant practice of the referring to SER as a percent, it must be understood that it is possible to have SER exceed 100%.

2.7.5 Digit Error Rate

Digit error rate (DER) is also a similar to WER, it is calculated in an identical manner. In some application, it can be difficult to achieve high digit accuracies. Levy, Linares, Bonastre, Stepmind, and Cannet (2005) reported their result DER is around 10.9%.

2.8 Summary

In this chapter, the literature review has been conducted to discuss many topics that relevant to acceptable accuracy using automatic transcription and phonetic labelling. The literature started with challenges for dyslexic children reading where while reading they tend to produce highly phonetically similar errors. So, the second aspect in this chapter focuses on overview of general architecture ASR system. The basic components of ASR in this chapter describe the functions of architectures that related with automatic transcription and phonetic labelling prior to performing training ASR engine.

The third aspect is about transcription and phonetic labelling using manual transcription and automatic transcription and phonetic labelling. Then, in this part there are three reviews of methods from automatic transcription and phonetic labelling namely forced alignment, neural network and morphological phonetic transcription. But, the commonly used by researchers of previous research is forced alignment method.

The fourth aspect is training using hybrid HMM/ANN prior evaluation accuracy of ASR. The review has been done on hybrid HMM/ANN because of manual and automatic transcription and phonetic labelling files were being train to construct ASR engine together with produce accuracy of ASR for evaluation. Then, the last aspect is evaluation accuracy performance of ASR engine using manual transcription versus ASR engine using automatic transcription and phonetic labelling. There are two standard metrics used for evaluating ASR engine which are WER and FAR.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

The methodology has been designed based on an understanding of literatures reviewed in Chapter 2 which relates to the objectives. Research methodology is a set of procedures or methods to explain why and how this research is intended to be conducted. The methodology comprises of four phases. The first phase is data collection of dyslexic children read speech in BM. The second phase is transcriptions of speech files using manual transcription and automatic transcription and phonetic labelling. After completing 585 speech files through transcription and phonetic labelling of both approaches, then in third phase, is training process using hybrid HMM/ANN. The training process was done separately using manual transcription files and automatic transcription and phonetic labelling files. Thus, in this study the training produced two ASR engines using different approaches of transcription and phonetic labelling that need to evaluate it accuracy performances. The last phase is evaluation of ASR engine accuracy using standard metric, WER and FAR. The overview of the methodology includes detail of each phase, activities, and methods as depicted in Figure 3.1.

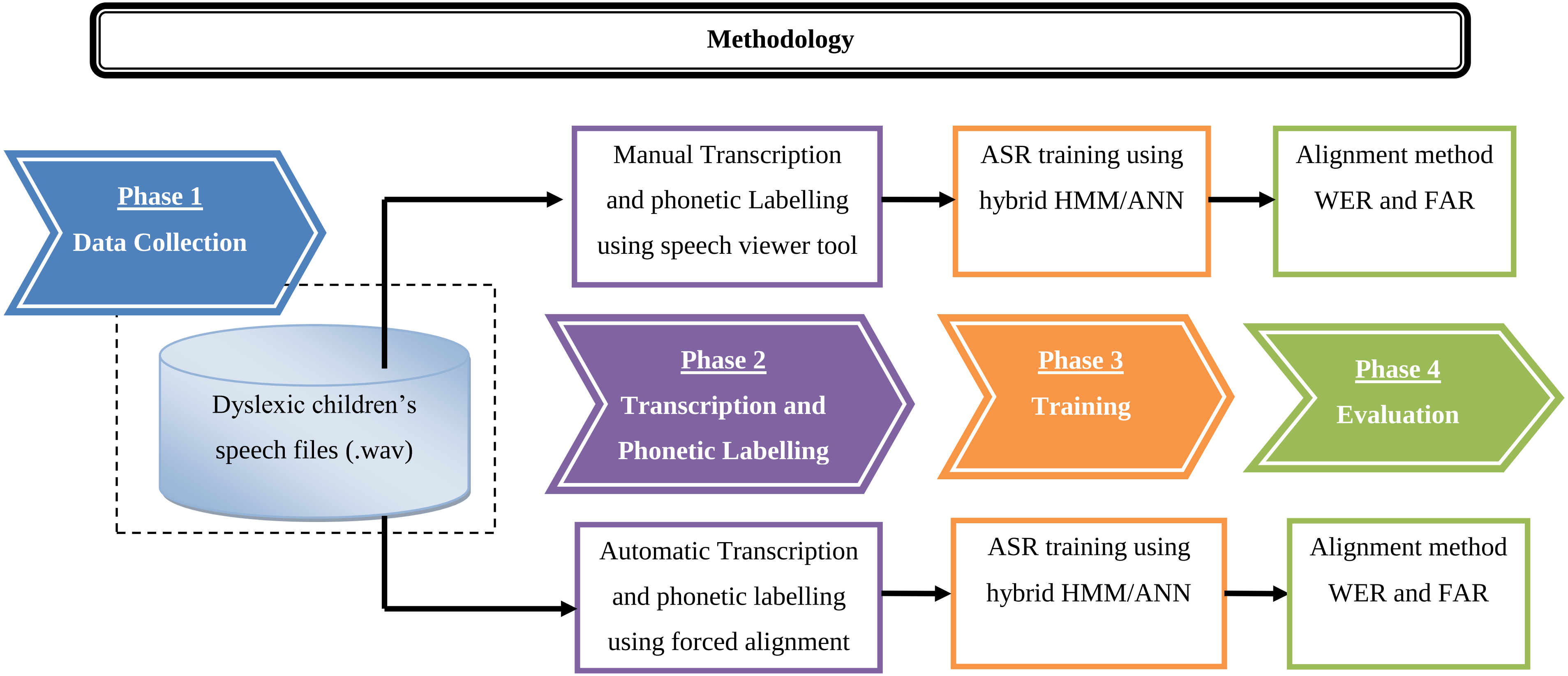

Methodology
Phase 1
Data Collection
Dyslexic children's speech files (.wav)
Manual Transcription and phonetic Labelling using speech viewer tool
ASR training using hybrid HMM/ANN
Alignment method WER and FAR
Phase 2
Transcription and Phonetic Labelling
Phase 3
Training
Phase 4
Evaluation
Automatic Transcription and phonetic labelling using forced alignment
ASR training using hybrid HMM/ANN
Alignment method WER and FAR

Figure 3.1. The Methodology.

3.2 Data Collection

Today, there are various efforts have been done to overcome learning disabilities of children with dyslexia in improve their reading, spelling and writing skill. Programs have been done to help dyslexic children's such as phonological awareness training Castles, Wilson, & Coltheart (2011), structured and multi-sensory approach McIntryre & Pickering (1995) and Davis Dyslexia correction program Gianna, Mclaughlin, Derby & Waco (2012). This available program can help dyslexic to overcome or reduce their reading difficulties. Besides, computer-based application also can be used as the fast track and efficient methods where dyslexic children would be fun using this application. The computer-based applications as seen as having to help dyslexic children such as Colorado Literacy Tutor (Colit) by Wise, Cole, Van, Schwartz, Snyder, Ngampatipatpong, et al. (2005) and STAR system by Russell et al. (1996). However, these programs are based on ASR technology that usually was built for English language and not suitable for Malaysian dyslexic children learning to read in BM.

For this study, secondary data have been used where the speech data were collected from a research conducted of speech recording of dyslexic children's reading aloud in BM (Husniza, 2010). The vocabulary for this study is 36 words reading aloud in BM that contain 585 speech files (.wav). The collection of speech data from ten children with dyslexia in primary schools, as young as 7 years old to 14 years old whose level reading is similar as identified and suggested by their teachers. Note that, the dyslexic children are allowed maximum two years extension of primary schooling in the Malaysian education system. This means dyslexic children age of 14 years old is still allowed to stay in the primary school if they are not ready for secondary school. The nature of children with dyslexia are difficulties to process the word, can't remember what words look like and skipping the letter during reading 36 words in BM causes most of speech files contain highly phonetically similar errors. According to Husniza (2010) the dyslexic children's speech was recorded using speech viewer tool of CSLU Toolkit (Sutton et al., 1998). The speech viewer function is to record every word that dyslexic children's read. A standard headphone with microphone is used in order to reduce noise from the environment. All these speech files were saved in .wav format to facilitate transcription and phonetic labelling process. In additionally, .wav files also use in training process together with .phn file and .txt file.

3.2.1 Data Description

The issues in this study about dyslexic children's reading contain highly phonetically similar errors that could affect accuracy of ASR. This section was analyst data collection in this work before commencing experiment for manual and automatic transcription and phonetic labelling and also training are conducted. In this work, there are 39 phonemes out of all 36 words in BM. Normally 39 phonemes from vowel, consonant, digraph, and diphthongs (e.g. ai, au, kh, kc). For BM syllable structure exist in the form CV, CVC, CVV and CVCC. Most of syllables are in the form of CV and CVC. According to Husniza (2010) each syllable pattern is present by taken randomly of syllabus in *Buku Panduan Pelaksanaan Program Pemulihan Khas (Masalah Penguasaan 3M)* for level one. Table 3.1 shows the different syllable of 36 words in BM.

Table 3.1. Different syllable pattern of 36 words in BM.

Syllable pattern	**Word**	**Syllable pattern**	**word**
V + CVC	abah	CV+CV+CV	kelapa
V + CVCC	abang	CV+CV+CVV	kemarau
V + CV	apa	CVC+CV+CVC	kumpulan
V + CV	aku	CVC+CV	makna
V + CVC	ayat	CVC+CV+CVC	maklumat
CV+ CV	baca	CV+CV with digraph	nyata
CV+CVCC	barang	V+CVCC	orang
CV + CVCC	bawang	CVC+CVC with diphthong	pandai
CV+CV+CVCC	belalang	CVCC+CV	pangsa
CV+CVC	betul	CVC+CV+CVCC	pendatang
CVC+CVC	bunga	CVC+CV	pergi
CVC+CVC	cantik	CVC+CVC	pernah
CVC+CV+CVC	cendawan	CV+CVCC	sayang
CV+CVV with diphthong	ceria	CV+CV+CV	selesa
CVC+CVC	hampar	CV+CV	suka
CVC+CV	jangan	V+CVCC	udang
CVCC+CV+CVC	jangkitan	V+CVC	umur
CV+CVC+CVCC	kecundang	CVC	wad

However, in this data collection there many errors made by nature dyslexic children’s reading due to condition that impede phonological awareness. A few types of errors made by dyslexic children’s reading in BM for this study like substitute vowel and consonant, Omits consonants or vowel and reverse the word.

According to Sawyer, Wade, and Kim (1999) errors made are grouped into corresponding category. Since the recordings were using dyslexic children’s reading the speech files involved are bounded by the most frequent error patterns. The type of errors involves in this study such as Substitutes Consonants, Substitutes Vowel, Omits Consonants, Omits Vowels, Reversals, Syllable Division confusion, Add consonants and Add Vowels. The most frequent error obtained in this study is omitted consonants when children read single word in BM. This error happen when dyslexic children deleted the consonant of word. For example, “jankitan” for *jangkitan*, “makumat” for *maklumat*, “pendatan” for *pendatang* and “hapar” for *hampar*.

Due to the phonological deficit, dyslexic children have normally somewhat different reading pattern. Children with dyslexia when reading tend to produce a lot of reading mistakes that sometimes are not obvious in normal children. Their symptoms of reading is can’t distinguished of similar shaped letters such as ‘b’ for ‘d’ or ‘p’ and vice versa, ‘m’ for ‘w’ and vice versa, ‘u’ for ‘n’ and vice versa. Hence, because of this the substitutions errors as found the second most commonly error in this work. There are two type substitutions which are substitutions consonants and substitutions vowels. For examples errors in substitutions consonants for word *belalang* replaced by *pelalang*, *jangkitan* instead *jangkidan*, *pangsa* instead *mangsa*, *jangan* instead *jangat* and *pergi* instead *bergi*. Then, for example errors substitution vowel in this study such as “ape” for *apa*, “kelape” for *kelapa*, “hampir” for *hampar*, and “urang” for *orang*.

The third most frequent errors in this study which are add consonant and add vowels. The example words were found in data collection that involved errors in add consonant and add vowel like “jangkitang” for *jangkitan*, “pendatangan” for *pendatang*, “kecungdang” for *kecundang*, “seleksa” for *selasa* and “peranah” for *pernah*. The others error obtained in this study is reversal error due to face for words that look similar. For example, “apah” for *abah*, “makau” for *makan*, and “wap” for *wad*.

All the error produced correspondence to selected pattern was also included in the active lexicon. The purpose of modelling using data collection that contains highly phonetically errors of dyslexic children’s because only the most frequent errors was to enable the development of ASR application that would be able to assist more dyslexic children. In this research 585 speech files has been used to completed automatic transcription and phonetic labelling using forced alignment and with similar amount of 585 speech files were perform using manual transcription. The phonetically similar error involved in this study is 20% of 585 speech files.

3.3 Transcription and Phonetic Labelling

The task of transcription and phonetic labelling is to transform speech files into small units named as phonetic symbol. In this study, the transcription and phonetic labelling considered two techniques which are manual transcription and automatic transcription and phonetic labelling. Both of these techniques produce 585 .phn files of 585 dyslexic children's read speech in BM. The 585 speech files are selected randomly from Husniza (2010). In this study, the transcription and phonetic labelling task begins with manual transcription as it is time consuming.

3.3.1 Manual Transcription

Manual transcription refers to the process whereby speech files perform transcription and phonetic labelling using human transcribers manually, referring to the spectrogram. Thus, there is no automatic assistance in transcription and phonetic labelling. In this study, the manual transcription act as benchmark for investigating the acceptable accuracy automatic transcription and phonetic labelling. This is because researchers believed manual transcription method is more accurate due to the use of human transcribers that ensures that transcription and phonetic label is perceptually valid (Dupuis, 2011; Dinarelli et al., 2009; Hazen, 2006; Gibbon, 1997). Furthermore, manual transcription requires human transcribers to hear sound of each phoneme of word prior to performing transcription and phonetic labelling.

In this study, the manual transcription and phonetic labelling took three months to accomplished 585 dyslexic children's speech files due to speech files contains highly phonetically similar errors. The highly phonetically similar errors of dyslexic children's speech are confusing to distinguish the phoneme of word. For examples, listening to the similar words *baca* and 'bace' or *wad* and 'wak'. Thus, it took a lot of time to accomplish each dyslexic children's read speech files.

The Worldbet symbols by Hieronymus (1993) was adopted as phonetic symbol for 36 words in BM. A Worldbet symbol is an American Standard Code for Information Interchanged (ASCII) representation based on the International Phonetic Alphabet (IPA). As shown in Table 3.2 the example Worldbet symbols in BM words.

Table 3.2. The example Worldbet symbols in BM words.

Words	**Worldbet**
Abah	`A (bc bh)A h`
Bawang	`bc bh A w A N`
Cantik	`tS A n tc th E kc k`
Jangan	`dZ A n A n`
Kecundang	`kc kh & tS U n dc d A N`
Kelapa	`kc kh & l A pc ph A`
Pangsa	`pc ph A N s A`
sayang	`s A j A N`
Suka	`s U kc kh A`
Wad	`w A (dc d)`

The "Speech Viewer" tool is used to perform manual transcription and phonetic labelling shown in Figure 3.2, which displays waveform of speech file. The speech viewer is a useful tool for laboratories designed to exhibit specific speech properties and it allows listening to the part of the speech signal corresponding with any phonetic symbol (Serridge, 2014).

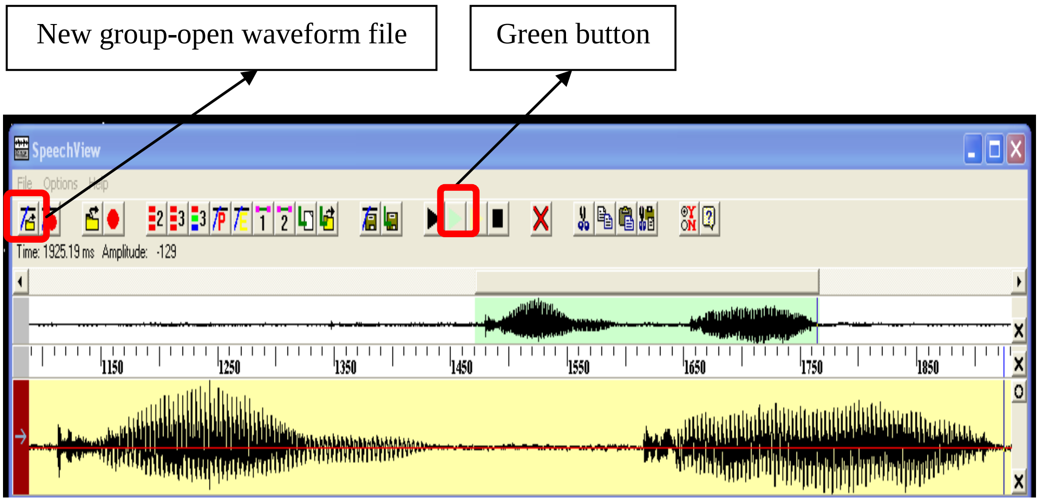


Figure 3.2. Speech view screen shot.

For comprehending process of manual transcription and phonetic labelling there are steps involved in order to transform dyslexic children's speech signal into transcription and phonetic labelling files:

i) Firstly, to upload speech signal in .wav format need to click on "new group-open waveform file" button in the toolbar as shown in Figure 3.2. This button is to uploaded speech files and automatically showed waveform.

ii) Secondly, to make sure that only the desired word (e.g. cantik) is included in the mapped region. The waveform should highlight from beginning of the word until the end of word as depicted in Figure 3.2. Then, green button is clicked to hear the sound of word we have selected is correct or not. In this task only word cantik should we highlighted, except of this word we need to label as a .pau means garbage from noisy environment or mistake that done by children dyslexic (e.g. "*urm", "aaa", "eh"* and etc).

iii) Now, the toolbar icon called "ADD-WINDOW: Gray-level 2-D spectrogram" or "ADD-WINDOW: Colour-level 3-D spectrogram" is clicked to view spectrogram. Figure 3.3 shown two type spectrogram of speech viewer; Gray-level 2D spectrogram, and Colour-level 3D spectrogram. Normally, in the transcription and phonetic labelling we choose either one; Gray-level 2D spectrogram or Colour-level 3D spectrogram. Then, to delete Grey-level 2D spectrogram or Colour-level 3D spectrogram we can click black letter "X" at the bottom right side. If you have removed both the spectrogram parameter dialog box for the gray-scale spectrogram window or colour-scale spectrogram window, bring it back by clicking on the black "O" options button.

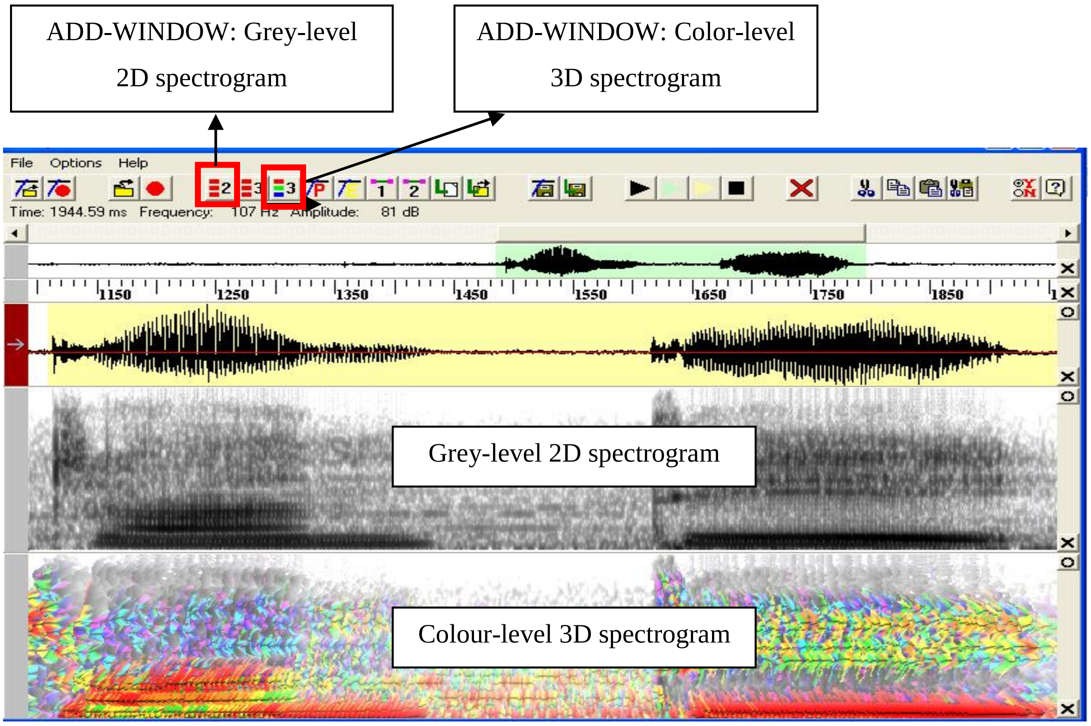
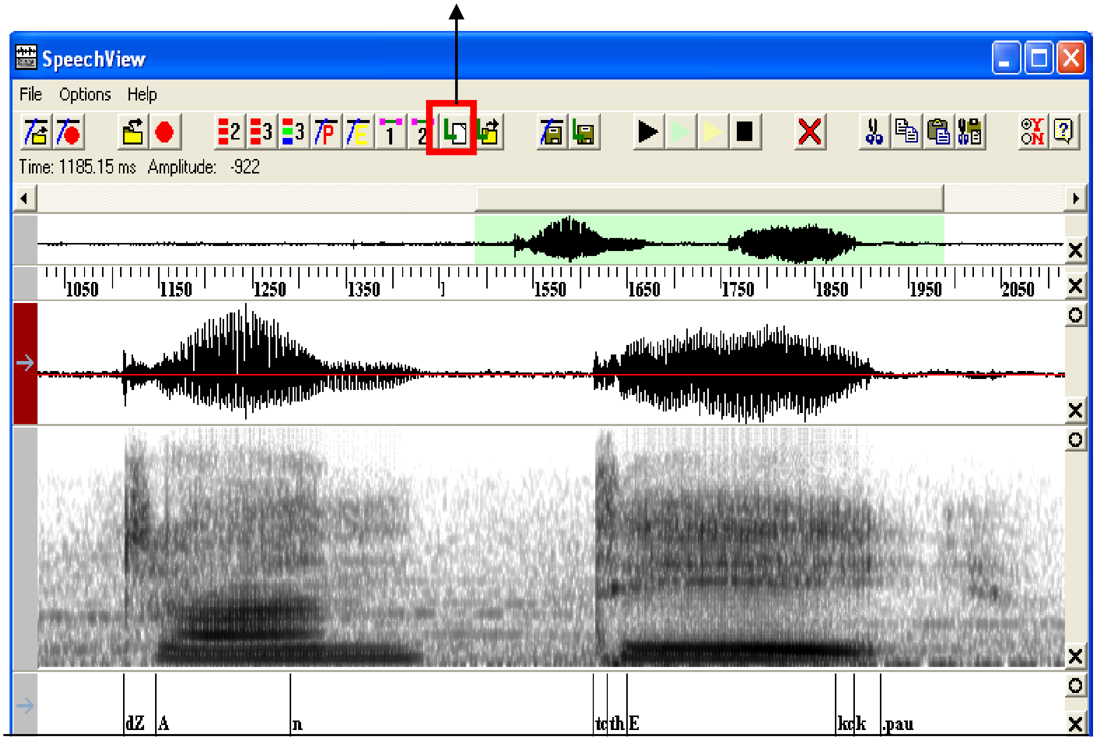

ADD-WINDOW: Grey-level 2D spectrogram
ADD-WINDOW: Color-level 3D spectrogram
File Options Help
Time: 1944.59 ms Frequency: 107 Hz Amplitude: 81 dB
1150
1250
1350
1450
1550
1650
1750
1850
Grey-level 2D spectrogram
Colour-level 3D spectrogram

Figure 3.3. Spectrograms of CSLU toolkit.

iv) Then, the transcription and phonetic labelling is started after clicked the toolbar icon labelled “Add Window: New Label Window” as shown in Figure 3.4. Use SHIFT plus the right mouse button clicks to create transcription boundaries and at the same time write phonetic symbols using Worlbet phoneme names “dZ A n tc th E kc k”. The phonetic symbol must be inserted one by one based on the sound of phoneme. The overall process manufacturing for transcription and phonetic labelling using manual transcription task is depicted in below in Figure 3.4 and for phonetic symbol in Figure 3.5.

Add Window: New Label Window”.



Time: 1185.15 ms Amplitude: -922

Figure 3.4. Manipulating the waveform, spectrogram and phonetic symbols associated with phonemes of word for manual transcription.

v) Lastly, click on the toolbar icon labelled "SAVE FILE: Label" in Figure 3.5 to save transcription and phonetic labelling file. The file must be saved in .phn format with the same name with speech files we choose (e.g. the speech files we select DC-1cantik1.wav then for transcription and phonetic labelling file were save as DC-1.cantik1.phn).

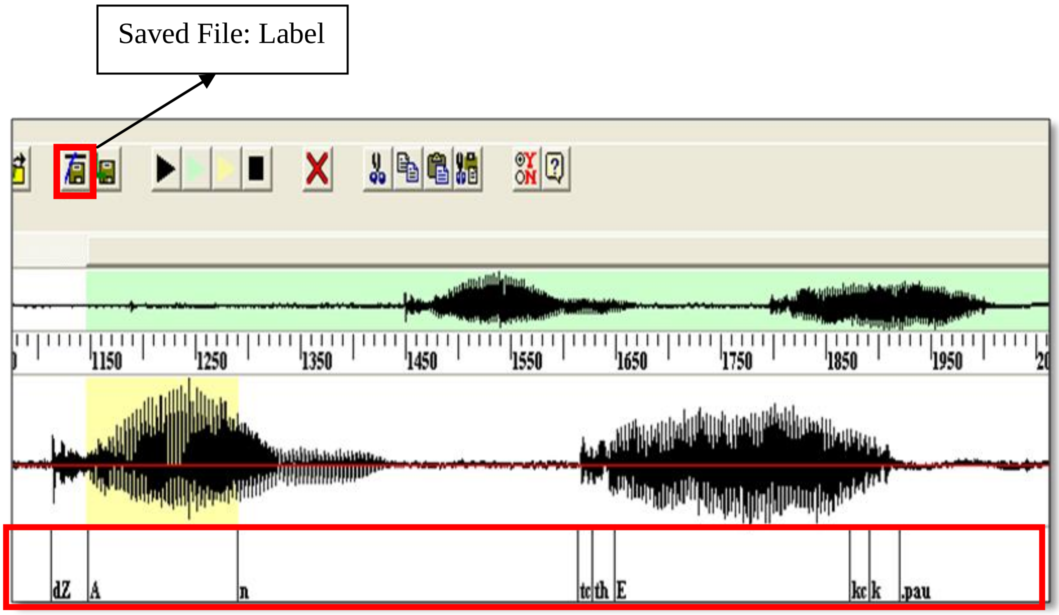


Figure 3.5. Phonetic symbols of word *“cantik”* and the speech signal that highlighted in yellow is phoneme for A.

For this study 585 dyslexic children’s speech files were accomplished where most frequent contain highly phonetically similar errors. A challenge of it cause for listen many times due to confusing sounds while they are reading aloud in BM. For example, listening the similar word like “*apah*” and abah. Therefore, this activity took three months to produce 585 .phn files. Then, the other technique for transcription and phonetic labelling is using forced alignment that done automatically.

3.3.2 Automatic Transcription and Phonetic Labelling

The availability of automatic transcription and phonetic labelling using forced alignment enable for us to present 585 dyslexic children's speech files faster rather than manual approach. Only two weeks required to accomplish transcribe, segment and label 585 speech files using automatic transcription and phonetic labelling. CSLU Toolkit (Sutton et al., 1998) is used to perform this task using forced alignment. The CSLU Toolkit is chosen in this study because it comes from complete packages to present manual transcription (speech viewer), automatic transcription and phonetic labelling (forced alignment) and also training ASR engine (HMM-ANN). The 585 .txt file must been created prior to performing automatic transcription and phonetic labelling using forced alignment. This files was created manually by hand that contain word depends of dyslexic children's reading aloud in BM. The other important files besides .txt prior to performing forced alignments are dyslexic children's speech files (.wav), lexicon file, spec file and neural network file. The description of all these files was illustrated in Table 3.3.

Table 3.3. Description five input files type prior forced alignment.

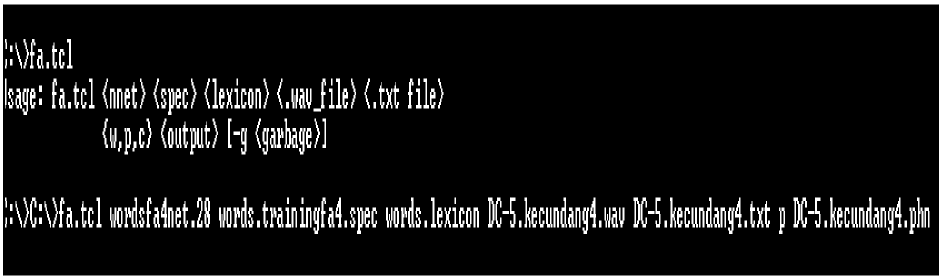
File type	**Description**
Speech file (.wav)	Contain read speech that recorded from the children reading aloud a single word in BM using speech view. The file format in waveform.

Text file (.txt)	Contain a text file according to what dyslexic children read.
Spec file (.spec)	Is the recognizer specification file.
Lexicon file	Contains a master list of each vocabulary and the location and format of the files for 36 words.
Neural network file (nntrain.exe)	The neural network files are generated by nntrain.exe. These files containing neural network weights to be used during phoneme classification.

As mention in Chapter 1, prior to forced alignment process the neural network file (nntrain.exe) and spec file which done in training using manual transcription files are required. In this study, the nntrain.exe file and spec file employ the one from Husniza (2010) due to similar speech features for 36 BM words and enable the Viterbi algorithm estimates of the probability the suitable phonetic symbols and locations for all boundaries of dyslexic children's speech files. Then, for spec file it contain the specific frame size, sampling rate, the location of code used to compute acoustic features, the context clusters, and phonetic mappings.

Hence, after all input files such as speech files (.wav), text file (.txt), lexicon file, spec file and neural network (nntrain.exe) are available then the command can be run to produced automatic transcription and phonetic labelling, which is .phn format.

The .phn is auto generated using forced alignment process. The *fa.tcl* was used to forced align 585 speech signal into transcription and phonetic symbols.



```
:\>fa.tcl
sage: fa.tcl <nnet> <spec> <lexicon> <.wav_file> <.txt file>
          <w,p,c> <output> [-g <garbage>]

:\>C:\>fa.tcl wordsfa4net.28 words.trainingfa4.spec words.lexicon DC-5.kecundang4.wav DC-5.kecundang4.txt p DC-5.kecundang4.phn
```

Figure 3.6. Command of forced alignment in produced automatic transcription and phonetic labelling.

After running fa.tcl command the system will read input files either it available or not. If the system can’t found location of speech files we need to run fa.tcl command again until fa.tcl can read all input files. Then, the nntrain.exe plays important role to search matching feature our dyslexic children’s speech (.wav) with its features in term of phonemes, phonetic symbols and all boundaries. Together with neural network (nntrain.exe) is lexicon file to help forced alignment find suitable phonetic symbols of each word. The lexicon file contains phonetic symbols of 36 words in BM. Refer Figure 3.7 for overall process of automatic transcription and phonetic labelling using forced alignment.

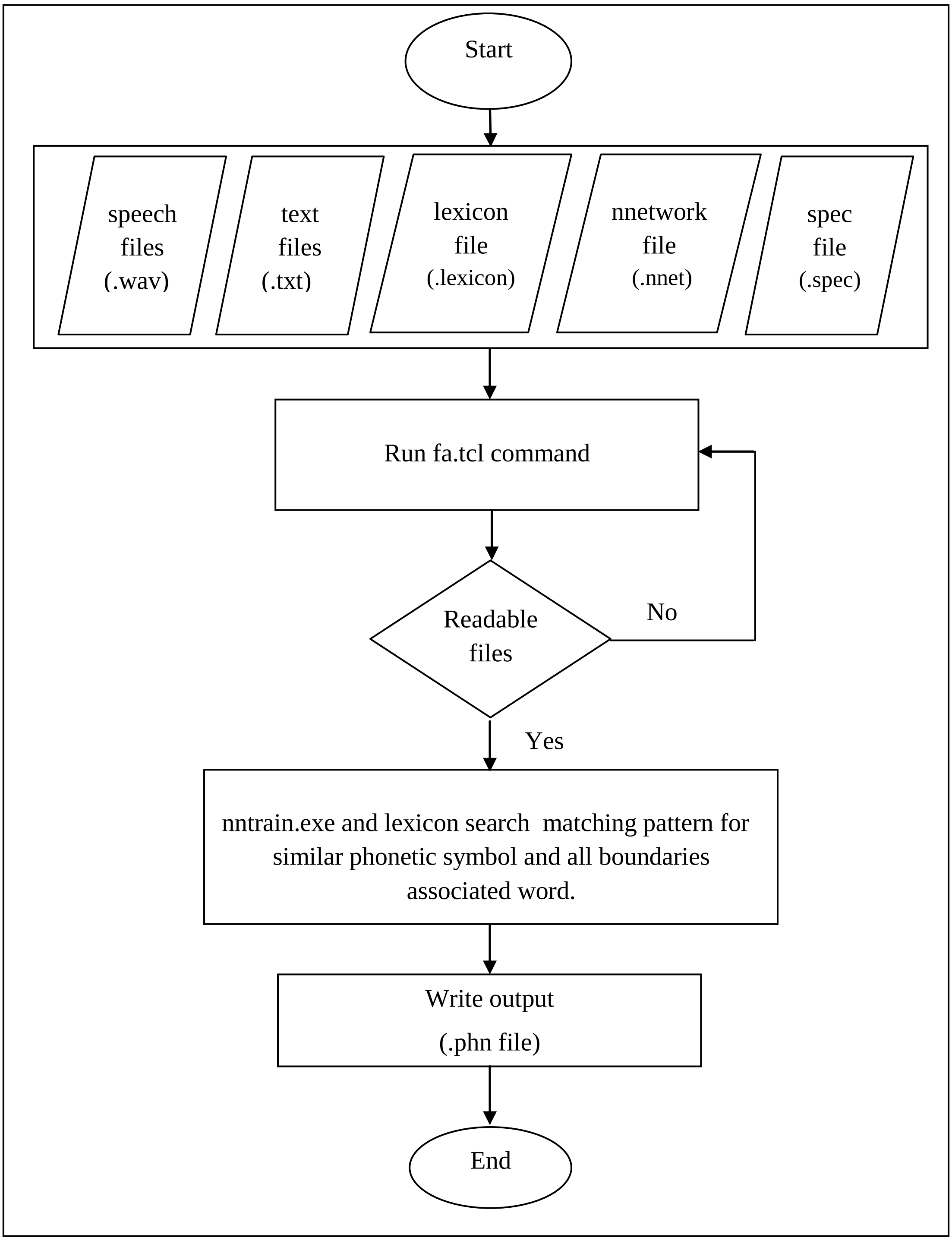
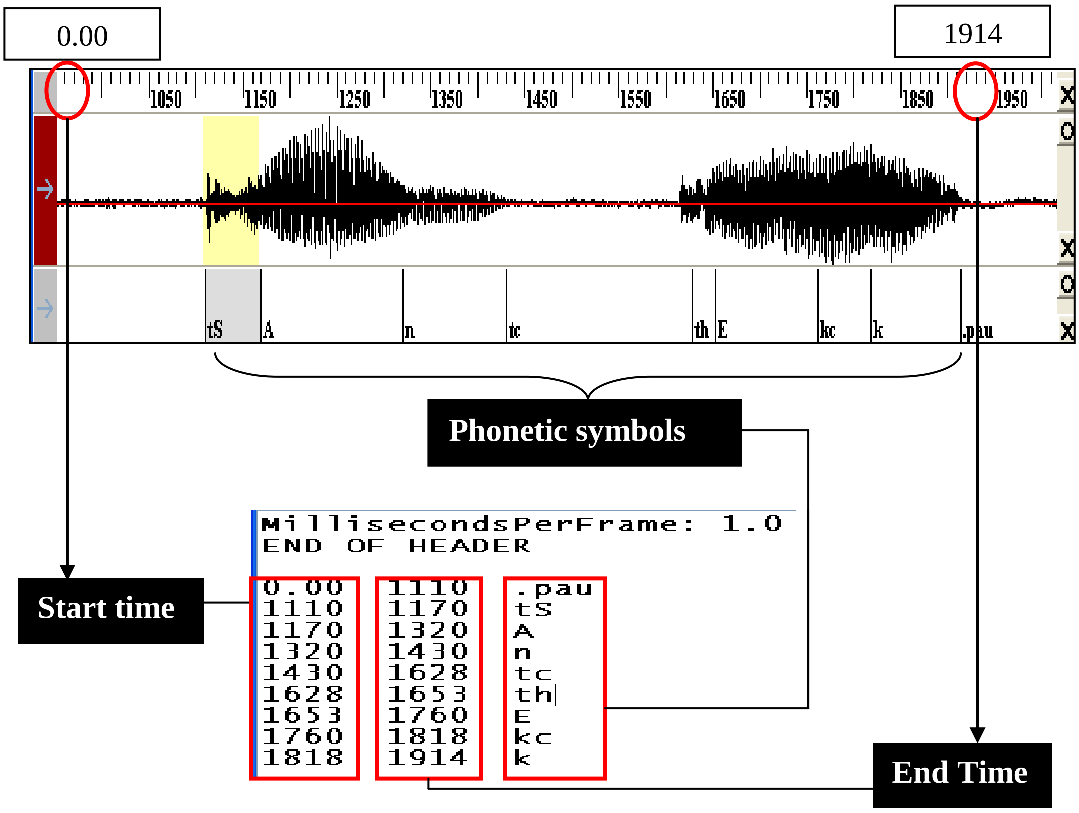


Figure 3.7. Process of automatic transcription and phonetic labelling using forced alignment.

The output file for automatic transcription and phonetic labelling is phoneme files in .phn format. Figure 3.8 shows the representation output of automatic transcription and phonetic labelling through forced alignment method. Every .phn files, has their own time duration which has been labelled with phonemes, as well as the starting time and ending time of each phoneme. There are three columns for each .phn file where first column is start time in milliseconds (ms), the second column is the end time in milliseconds (ms), and the third column is the phonetic symbols for that segment. The first two lines of the file are a header which defines the length of a "frame" in milliseconds (ms). The rest of the files consist of two numbers that define a frame range, and a label that applies to that region. For example:



```
MillisecondsPerFrame: 1.0
END OF HEADER
0.00  1110  .pau
1110  1170  tS
1170  1320  A
1320  1430  n
1430  1628  tc
1628  1653  th
1653  1760  E
1760  1818  kc
1818  1914  k
```

Figure 3.8. Automatic transcription and phonetic labelling for the word *"cantik"*.

So, as can seen in Figure 3.8 shows that a frame corresponds to 1 millisecond (ms) of time, and that from 0.000 to 1110.0 ms into the file, there is a pause (.pau), with the first phoneme symbol of word "*cantik"* is `(tS)` starting at 1110 ms and stretching to 1170 ms. Then, for last boundary of word "*cantik*" is phonetic symbol `(k)` their start time starting is 1818 and end time to 1914 for its boundary.

After automatic transcription and phonetic labelling is completed, the comparative accuracy between transcription using automatic and manual approach need to be examined. Cosi and Hosom (1991) state that "the accuracy of automatic transcription and segmentation need always to be checked using references manual transcription by phonetic or speech communication". Thus, both techniques transcription and phonetic labelling files need to be trained for evaluation using WER and FAR.

3.4 Training using Hybrid HMM/ANN

In order to investigate acceptable accuracy of automatic transcription and phonetic labelling using forced alignment, this study produces ASR engine using manual transcription and ASR engine using automatic transcription and phonetic labelling. The training was done by using hybrid HMM/ANN also of CSLU toolkit (Hosom, 2006) and it is available to download and used for research purpose only. Additionally, the hybrid HMM/ANN approach of CSLU toolkit has been trained with children's read speech which gives essential significant in this study.

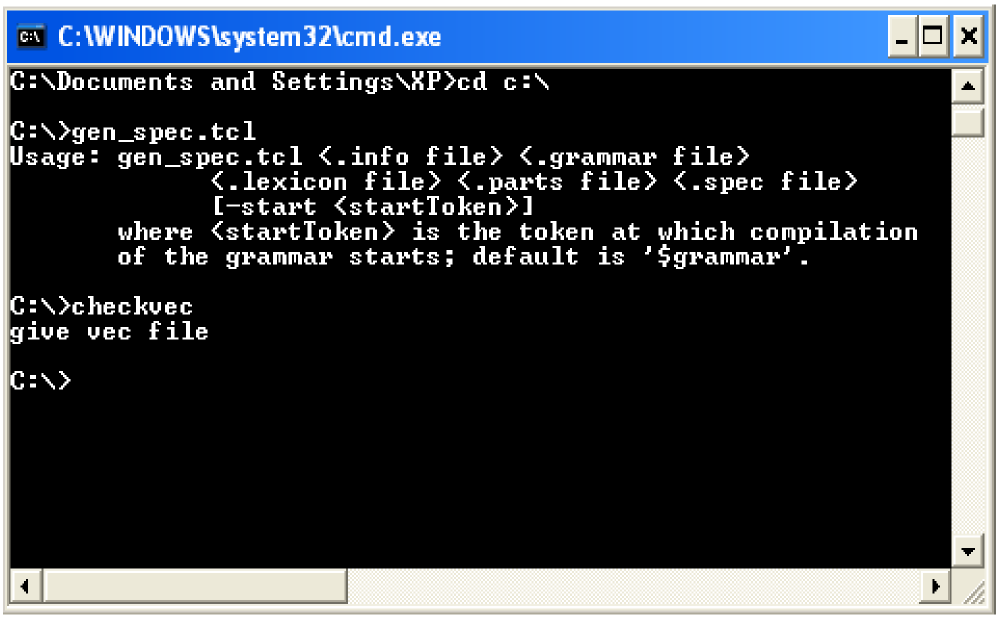
This hybrid HMM/ANN training process consists of executing a sequence of CSLUsh using description files that specify aspects of the training condition, corpora, and the recognizer architecture. CSLUsh is the main application level of the toolkit and combination of the core technology modules with the well known easy-to-learn Tcl/Tk scripting language and freely available to download. In order to develop ASR engine using the constructed automatic transcription and phonetic labelling and also manual transcription, the description files are created manually and several CSLUsh scripts are used to perform training of an ASR engine. There are five main steps to creating an ASR engine through hybrid HMM/ANN training which are setting directory, create description files, find data for training, start training to develop ASR engine, and retrain to get ASR engine with best accuracy. The training process requires several iteration using manual transcription files and automatic transcription and phonetic labelling files to get the best accuracy results of ASR engine. In this study, both transcription and phonetic labelling go through the same process of training that has been illustrated in the following section.

3.4.1 Setting directory

Firstly, before training process, manage or setup path in our computers to make sure any of the toolkit's command during training process working properly. The training process uses commands entered from a DOS prompt[1], a DOS window can be found

[1] A command prompt is used in a text-based or "command-line" interface, such as a Unix terminal or a DOS shell. It is a symbol or series of characters at the beginning of a line that indicates the system is ready to receive input.

from Start » Programs » Accessories » Command Prompt. A command window can be added to the start menu for easy access. Figure 3.9 depicts example command prompt to set prior training process.



```
C:\Documents and Settings\XP>cd c:\

C:\>gen_spec.tcl
Usage: gen_spec.tcl <.info file> <.grammar file>
             <.lexicon file> <.parts file> <.spec file>
             [-start <startToken>]
       where <startToken> is the token at which compilation
       of the grammar starts; default is '$grammar'.

C:\>checkvec
give vec file

C:\>
```

Figure 3.9. Example command prompt used in training an ASR.

Besides setting the directories, the 585 speech files (.wav) gathered from data collection phase, the transcription and phonetic labelling of each approach that done manually and automatically in .phn format and also text file (.txt) are required before the training process begins. Figure 3.10 illustrates relationship between speech file (.wav), text file (.txt) and transcription and phonetic labelling file (.phn). After these three files are available then proceed to follow instruction by Hosom (2006) to create description files and begin training an ASR engine.

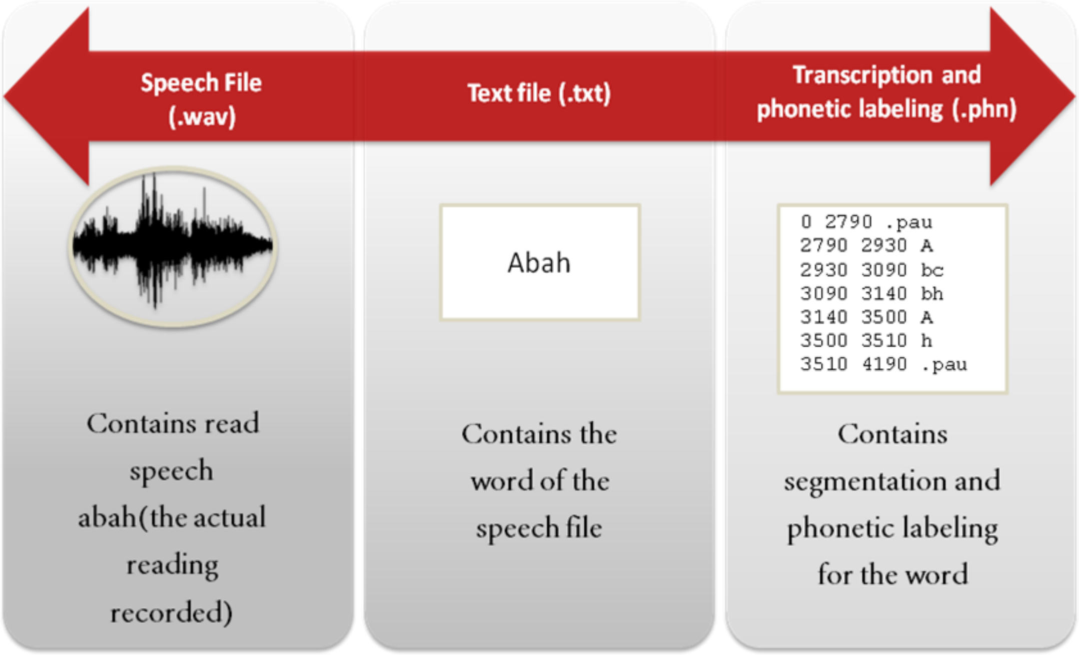

Speech File (.wav)
Text file (.txt)
Transcription and phonetic labeling (.phn)
Abah
0 2790 .pau
2790 2930 A
2930 3090 bc
3090 3140 bh
3140 3500 A
3500 3510 h
3510 4190 .pau
Contains read speech abah(the actual reading recorded)
Contains the word of the speech file
Contains segmentation and phonetic labeling for the word

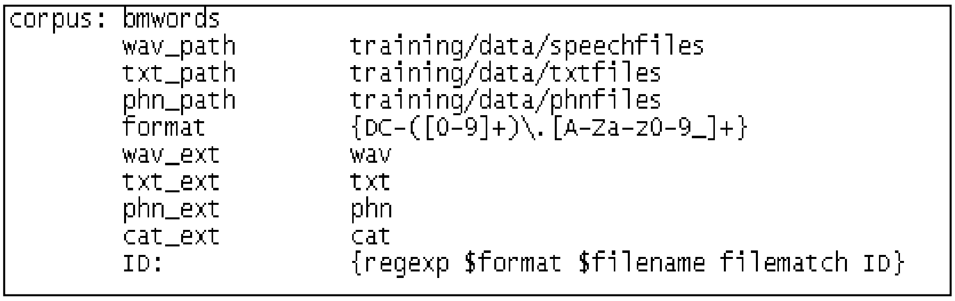
Figure 3.10. The relationship between speech files, text files and transcription and phonetic labelling files.

3.4.2 Create Description Files

The description files such as corpora.txt, an info file for each training, development, and testing data set, lexicon file and a part file was created manually. These files required to be mandatory for training.

i) Corpora.txt

The corpora file contains basic information and master list for each location of each files (.wav, .phn and .txt files). The name of file format is specified and it is compulsory to have the same name (e.g. DC-1.cendawan3.wav, DC-1.cendawan3.txt and DC-1.cendawan3.phn). There is no automated way of generating this file, but it is easy to modify by hand. The same corpora file can be used for all training tasks. Thus, the corpora.txt file used in this study is as written below:-



```
corpus: bmwords
        wav_path        training/data/speechfiles
        txt_path        training/data/txtfiles
        phn_path        training/data/phnfiles
        format          {DC-([0-9]+)\.[A-Za-z0-9_]+}
        wav_ext         wav
        txt_ext         txt
        phn_ext         phn
        cat_ext         cat
        ID:             {regexp $format $filename filematch ID}
```

Figure 3.11. The corpora file for the training process.

ii) Lexicon file

The lexicon file created to have specified the pronunciation of each word in the grammar. Lexicon files for this study models using 36 of BM words from dyslexic children’s reading aloud in BM. This lexicon file as known as lexicon model constructed with phoneme refinement and treat mispronunciations as alternative pronunciation for an improved accuracy. Figure 3.12 showed lexicon file in this study, which has been modified to improve accuracy of ASR engine.

```
abah = A(bc bh)|(pc ph) A h;
abang = A (bc bh)|(dc dh)|(pc ph) A N;
apa = (A (pc ph) A|&)|(bc bh A pc ph A);
aku = A kc kh U;
ayat = (A j A tc t|h)|(A j A);
baca = bc bh A|& tS|dZ A|&;
barang = bc bh A|I 9 A|I N;
bawang = bc bh A|& w A N;
belalang = bc bh & l A l A N;
betul = bc bh & tc th o l;
bunga = bc bh U N A;
cantik = tS|dZ A n tc th E kc k;
cendawan = tS|dZ &|I n dc d A w A n;
ceria = tS & 9 ia;
hampar = h A m pc ph A|I 9|(tc t);
jangan = dZ|tS A N A n;
jangkitan = dZ|tS A N|n (kc kh)|(gc g) I|E tc th A|I n;
kecundang = kc kh & tS|dZ U n dc d A N|n;
kelapa = kc kh & l A pc ph A;
kemarau = kc kh & m A 9 aU|aI|oU;
kumpulan =(kc kh U|& m pc ph U l A n)|(kc kh U pc ph U l
           A n);
makan = m A kc k A n;
maklumat = (m A kc k l U m A tc t)|(m A l U m A tc t);
nyata = n~ A tc th A;
orang = o 9 A N;
pandai = pc ph A n dc d aI;
pangsa = pc ph A N|n s A;
pendatang = (pc ph)|m & n dc d A tc th A N;
pergi = pc ph & 9 gc g I;
pernah = pc ph & 9 n|(tc th) A h;
sayang = s A j A N;
selesa = s & l E|A s A;
suka = s U kc kh A;
udang = U dc d A N;
umur = U m O 9;
wad = w A (dc d)|n;
```

Figure 3.12. Lexicon file for 36 BM words that has been ‘cleaned’ for better accuracy.

iii) Grammar file

Create a “grammar” file that specifies the grammar that will be used to recognize words. The grammar file contains the structure of each speech attribute in a .wav file. Thus, the grammar of each word level reading is easier to recognize. Hence, it is used in this study to follow that of a discrete recognition that gives $grammar = [*sil%%] $word [*sil%%] where [*sil%%] refers to garbage or silence and $word takes the word spoken. The grammar file for this study is depicted as below:

```
$word= abah | abang | apa | aku | ayat | baca |
       barang |  bawang | belalang | betul | bunga|
       cantik | cendawan | ceria | hampar | jangan |
       jangkitan | kecundang | kelapa | kemarau |
       kumpulan | makan | maklumat | nyata | orang |
       pandai | pangsa | pendatang | pergi | pernah
       | sayang | selesa | suka | udang | umur |
       wad;

$grammar = [*sil%%] $word [*sil%%];
```

Figure 3.13. Grammar file that involved 36 vocabularies in BM.

iv) Info file

The .info file contains information on the training dataset, development dataset and testing dataset as separate files. Each info file describes corpus information for which datasets would be trained, developed, and tested. The corpus file contains all of the information that is important to find examples for training, development, or testing. The info files include 3 partitions which are training dataset, development dataset and testing dataset. The division follows the 3:1:1 ratio as required by the CSLU Toolkit. The training requires three dataset because of once train is done using training dataset the system were given the best accuracy of the best network. The best result of training dataset is used to train development dataset. In development data set the training using hybrid HMM/ANN were train several iterations until the result of developments dataset is reduce then the training process need to stop. Hence, after stop train development dataset, the training for test dataset as the final result for ASR were used the best network that given the best result to train test dataset. In this work, the result of test dataset as final results for ASR was used to measuring WER and FAR.

The files are automatically separated by running find_files.tcl. It is very important to make sure that the data in the testing set do not occur in the training and development datasets. The total of files for this 3 partition is 585 files. This dataset was used in order to evaluate word-level performances with a larger network and search parameters in a reasonable amount of time. Hence, a total of 357 speech files are used for training, 132 files for development, and 96 for testing, as depicted in Figure 3.14.

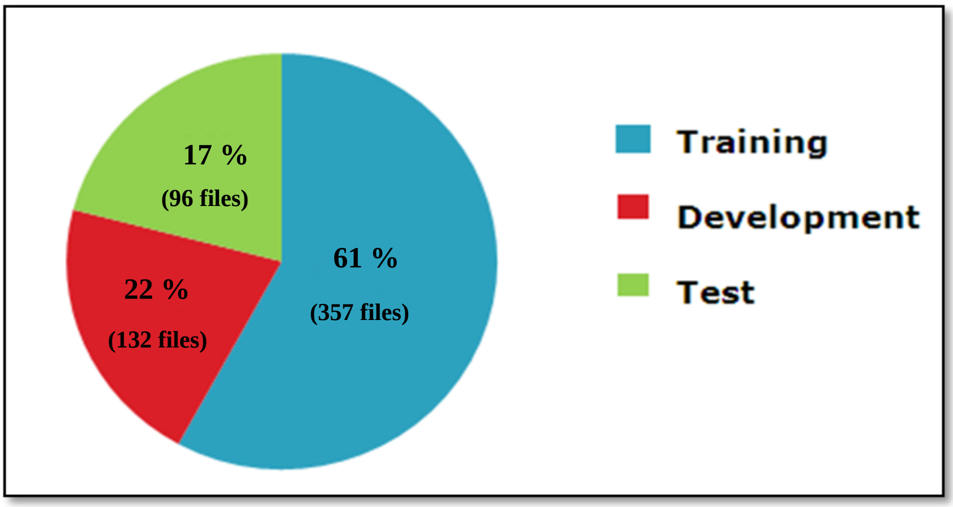
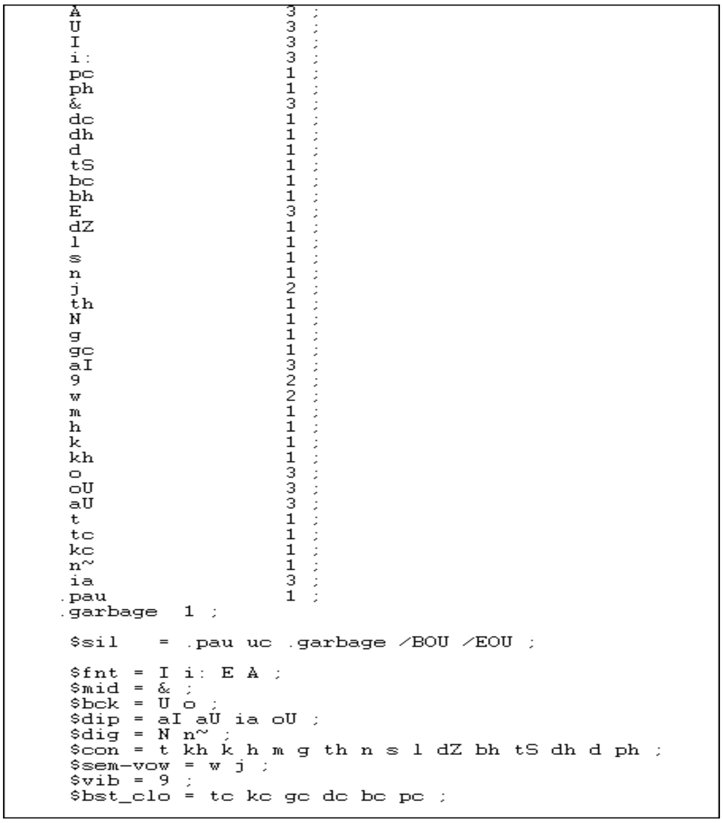


Figure 3.14. The files are automatically separated by running find_files.tcl.

v) Part files

The part specification can be referred to as the phonemes specification to determine whether a phoneme is context-dependent or context-independent. Context-dependent modelling is performed to allow for variation in the speech signal for the same phoneme to be regarded and trained (Zulikha & Husniza, 2010a). This part files contains a number of parts that each phoneme has been split and mapped from a one phoneme to another symbol. The parts file for each phoneme can be split into one part, two part or three part. Refer to Figure 3.15.



```
 A                    3 ;
 U                    3 ;
 I                    3 ;
 i:                   3 ;
 pc                   1 ;
 ph                   1 ;
 &                    3 ;
 dc                   1 ;
 dh                   1 ;
 d                    1 ;
 tS                   1 ;
 bc                   1 ;
 bh                   1 ;
 E                    3 ;
 dZ                   1 ;
 l                    1 ;
 s                    1 ;
 n                    1 ;
 j                    2 ;
 th                   1 ;
 N                    1 ;
 g                    1 ;
 gc                   1 ;
 aI                   3 ;
 9                    2 ;
 w                    2 ;
 m                    1 ;
 h                    1 ;
 k                    1 ;
 kh                   1 ;
 o                    3 ;
 oU                   3 ;
 aU                   3 ;
 t                    1 ;
 tc                   1 ;
 kc                   1 ;
 n~                   1 ;
 ia                   3 ;
.pau                  1 ;
.garbage  1 ;

 $sil    = .pau uc .garbage /BOU /EOU ;

 $fnt = I i: E A ;
 $mid = & ;
 $bck = U o ;
 $dip = aI aU ia oU ;
 $dig = N n~ ;
 $con = t kh k h m g th n s l dZ bh tS dh d ph ;
 $sem-vow = w j ;
 $vib = 9 ;
 $bst_clo = tc kc gc dc bc pc ;
```

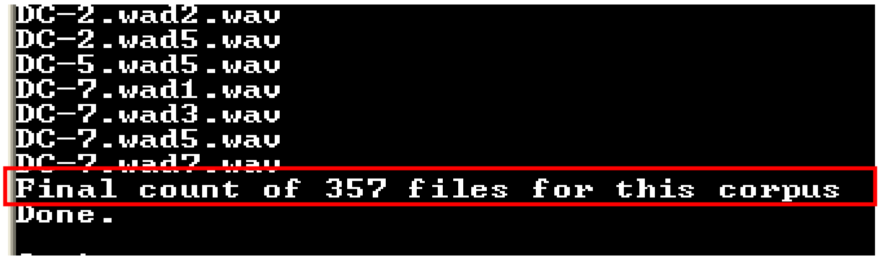
Figure 3.15. The parts file.

3.4.3 Find Data for Training

After we accomplish creating five files above then we proceed this phase by running some command to find data for training. There are following scripts was used for prior select data for training an ASR engine.

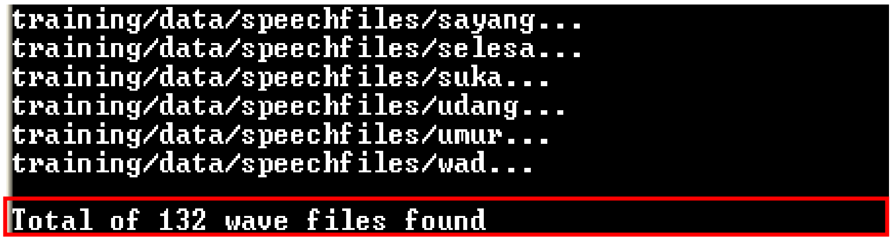
i) Find files.tcl

find files.tcl is to find files for training, development, and testing according .info files. When executed, this command gives a list in a particular dataset based on the requirement specified in the info files. The files for each dataset are determined by the specific ratio for the three datasets. As a suggestion from the CSLU toolkit tutorial, 3/5 file consists of training and 1/5 file for each development and testing dataset. Figure 3.16, Figure 3.17, and Figure 3.18, depict a snippet of the output generated of find files.tcl. for development, training, and test dataset. The files are automatically grouped into their corresponding dataset: Words.training.bmwords.files, words.dev.bmwords and words.testing.bmwords.files.



```
DC-2.wad2.wav
DC-2.wad5.wav
DC-5.wad5.wav
DC-7.wad1.wav
DC-7.wad3.wav
DC-7.wad5.wav
DC-7.wad7.wav
Final count of 357 files for this corpus
Done.
```

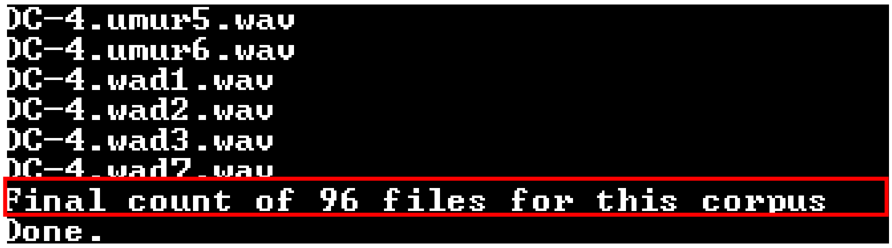
Figure 3.16. Number of files for training.



```
training/data/speechfiles/sayang...
training/data/speechfiles/selesa...
training/data/speechfiles/suka...
training/data/speechfiles/udang...
training/data/speechfiles/umur...
training/data/speechfiles/wad...

Total of 132 wave files found
```

Figure 3.17. Number of files for development.



```
DC-4.umur5.wav
DC-4.umur6.wav
DC-4.wad1.wav
DC-4.wad2.wav
DC-4.wad3.wav
DC-4.wad7.wav
Final count of 96 files for this corpus
Done.
```

Figure 3.18. Number of files for testing.

ii) gen_spec.tcl

gen_spec.tcl is to determine the context-dependent categories that were classified by the recognizer. To run this command we used files that already created before (e.g. info file, grammar file, lexicon file and part files). Additionally, the specification file contains the specific frame size, sampling rate, the location of code used to compute acoustic features, the context clusters, and any phonetic mappings.

iii) gen_catfiles.tcl

The command gen_catfiles.tcl is used to take the list of files for training (words. trainings. words. files) and create time-aligned categories from text transcription or from phonetic time-aligned transcriptions. These categories are written to separate files with the extension ".cat", which are put in sub directories that mirror the directory structure of the corpus being used. The outputs after running gen_catfiles.tcl are words.trainingfa.dur and words.trainingfa.counts.

iv) revise_spec.tcl

The command is used to make sure that have enough example of each category to training and additionally add duration limits to update minimum and maximum duration parameters for each category. Moreover, this command created output files that indicate the number of examples available for each category, as well as the duration information. The output of this script is modified "spec" file.

3.4.4 Select Data for Training

Once the files have been selected, the category files have been created and the description file is correct, then the following scripts and command to select frame for trainings are used:

i) pick_examples.tcl

pick_examples.tcl is used to select examples to train on. The examples file is an ASCII file that describes the location such as filename and frame number of each category that will be trained. The output of this script is an “examples” file, which is used directly by the next script, gen_example.tcl. Figure 3.19 shown the examples of speech files for training

```
/training/speechfiles/barang/MD-5.barang1.wav
5 33
5 174
19 96
...
/training/speechfiles/bawang/MD-1.bawang3.wav
5 98
32 82
19 51
....
```

Figure 3.19. The available example files for training.

ii) gen_examples.tcl

The command gen_examples.tcl is used to compute acoustic features for all of the frames given in words.trainingfa.examples of running command pick_examples.tcl. This command creates a binary file with the extension ".vec" (for vectors of features).

iii) checkvec.exe

Use checkvec.exe to make sure that the vector file that has been created has the correct format, and that every category has at least one example to train on. The numbers in the left column are the values corresponding to each category (from 1 to the total number of categories), and the numbers in the right column are the number of examples for each category.

3.4.5 Training ASR engine

Then after data for training is selected, the training ASR engine continues with run nntrain.exe and select_best.exe to get the accuracy rate for evaluation task. The training is not only done in one process, but need to retrain until the best accuracy is obtained or recognition rate shown reduction result. The results of training using manual transcription and automatic transcription and phonetic labelling were discussed in Chapter 4. These are the following command to produce recognition rate for ASR engine that involved node, input layer, hidden layer, output layer and learning rate.

i) nntrain.exe

The "nntrain" used to train the neural network iterations using the vector file `words.train.vec` as training data. This program creates a weights file at each iteration where the best weights file was selected after train 30 iterations. The 30 iterations are used in this work because of it is suggested in the documentation. However, the setting for iteration can be changed but must 30 and below iterations. A number of 130 inputs layer units and 200 hidden layer units are employed for the standard feature for CSLU toolkit. The number of the output layer units, usually depends on the total number categories that considered to be trained. Thus, in this case, 78 output layer units are used base on vector file created. Figure 3.20 the illustration of structure chart neural network.

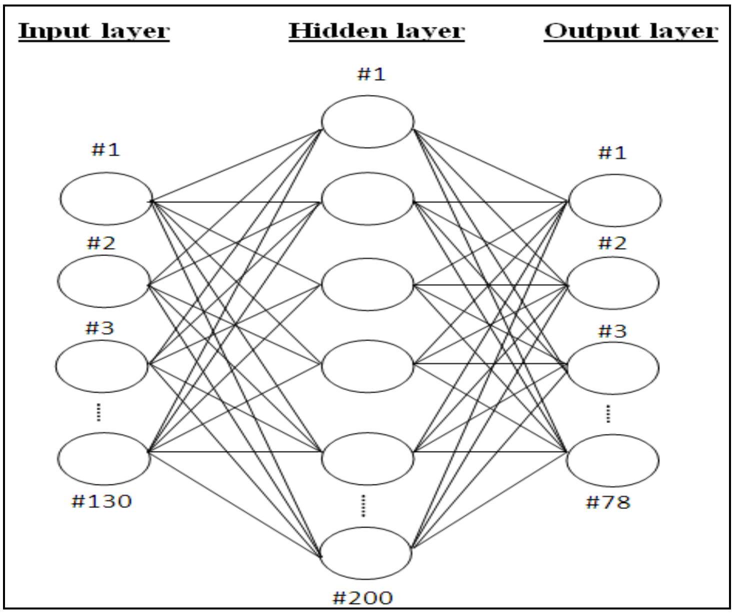


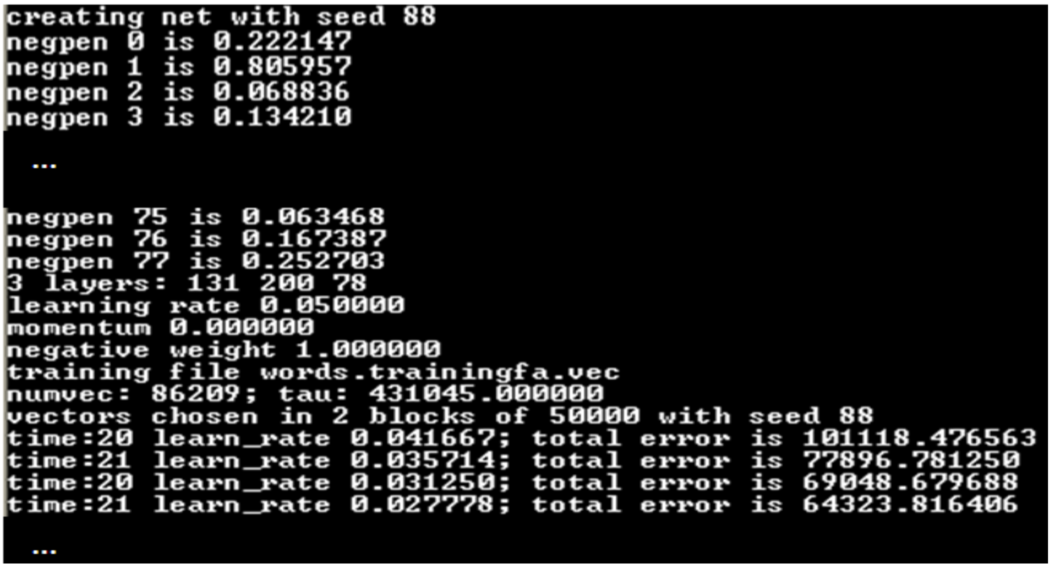
Figure 3.20. The structure chart network architecture.

The network is trained by execute nntrain.exe command of the toolkit to generated output layer. Its start weight is set to -1.0 which their learning rate set randomly at 0.05. The function of weight is to multiply the signal transmitted from input layer to hidden layer. Referring Table 3.4 the parameters before executed nntrain.exe command.

Table 3.4. The parameters in execute nntrain.exe command.

`-l`	Allow for negative penalty (negpen)
`-sn 88 –sv88`	Random seed value
`-f wordsnet`	basename for output weight file called ‘wordsnet’.
`-a 3 130 200 78 30`	Architecture: 3 layer, 130 input nodes, 200 hidden nodes, 78 output nodes with 30 iteration
`Words.train.vec`	Vector file

The learning rate is automatically adjusted to minimize the total error. The total error should reduce in every interaction. In this case, the error ratio obtained is 0.73, which is acceptable because according to Hosom (2009), the acceptable ration range from 0.5 until 0.9 where the standard ration is 0.75. Figure 3.21 shows the results for learn rate and total errors while execute nntrain.exe.



```
creating net with seed 88
negpen 0 is 0.222147
negpen 1 is 0.805957
negpen 2 is 0.068836
negpen 3 is 0.134210

  ...

negpen 75 is 0.063468
negpen 76 is 0.167387
negpen 77 is 0.252703
3 layers: 131 200 78
learning rate 0.050000
momentum 0.000000
negative weight 1.000000
training file words.trainingfa.vec
numvec: 86209; tau: 431045.000000
vectors chosen in 2 blocks of 50000 with seed 88
time:20 learn_rate 0.041667; total error is 101118.476563
time:21 learn_rate 0.035714; total error is 77896.781250
time:20 learn_rate 0.031250; total error is 69048.679688
time:21 learn_rate 0.027778; total error is 64323.816406

  ...
```

Figure 3.21. The result for learn rate and total errors while training the hybrid HMM/ANN.

ii) select_best.exe

The select_best.exe to evaluate the performance of each iteration (weight file) on the dataset. This command usually takes a long time, if there are many files in each dataset. In this work, the best iteration of the network is given by the best result word accuracy. The explanation best iteration and best result accuracy were discussed more in Chapter four.

3.5 Evaluation of ASR Accuracy

After training is done and best results are obtained using manual and automatic transcription and phonetic labelling then the accuracy of ASR engines must be evaluated. Evaluating accuracy of ASR engine was measured using WER and FAR. We choose WER as metrics because previous studies acknowledged that WER is a highly valid metric and it also commonly quoted measurement accuracy of ASR engine (Hagen et al., 2006; Russell et al., 1996). According to Jurafsky and Martin (2000) good performance accuracy of ASR is indicated by a lower percentage of WER.

The evaluation of accuracy is not enough if using WER alone because it provides only read speech independent of the target word or text (Mostow, 2006). So, we used FAR to evaluate its performance in terms of miscue detection. The FAR can support the ASR accuracy measurement while it recognizes dyslexic children reading error (Lee et al., 2004; Mostow, 2006). Both standard metrics WER and FAR were used to perform the evaluation of ASR accuracy (Mustafa et al., 2015; Hagen, 2006; Liu et al., 2008).

Thus, in this study two ASR engines using manual and automatic transcription and phonetic labelling have been developed. WER and FAR were used to evaluated each ASR engine because of aims this study to investigate acceptable accuracy for ASR engine using automatic transcription and phonetic labelling. The acceptable accuracy of ASR engine using automatic transcription and phonetic labelling is depends on result accuracy of ASR engine using manual transcription. As mentioned in Chapter One manual transcription is more accurate (Goldman, 2011; Kim & Gibbo, 2011; Mporas, Ganchev & Fakotakis, 2010; Kabir & Giurgiu 2012). The results accuracy of an ASR engines using automatic transcription and phonetic labelling is acceptable if it is at par with the accuracy of ASR engine using manual transcription.

For WER and FAR the lower percentage is the better accuracy for ASR engine. Then for FAR, the evaluation is about where a total of 128 speech files were randomly selected among available speech files (that were not used in training, development and testing datasets) to measure its miscue detection performance. Note that the selected files are of acceptable quality where they contain less background noise or garbage. To evaluate ASR engines formula that is shown below was used to calculate quality accuracy of both ASR engines. The detail results about this evaluation in phase four were discusses more in Chapter 4.

i) Formula of WER

$$\text{WER} = 100\% - \text{recognition rate } \%$$

ii) Formula of FAR

$$\text{FAR} = \frac{\text{Number of correct readings recognized as incorrect}}{\text{Total number of correct readings}} * 100$$

3.6 Summary

The methodology comprises four phases. The first phase is data collection. Data collection obtains the secondary data based on speech recordings of dyslexic children reading aloud certain words in BM. The second phase involved two techniques using manual technique and automatic technique. In this phase, the steps to perform manual transcription using speech viewer tool and automatic transcription and phonetic labelling using forced alignment has been elaborated and additionally Worldbet symbols were used as the phonetic symbols representation. After completing the work to produced 585 .phn files using manual transcription and 585.phn using automatic transcription and phonetic labelling. Then, the activities in this study proceed to train both transcription and phonetic labelling (.phn) files separately. The trainings in this study are done in five iterations to get satisfaction result and stop the training after results shown reduction of accuracy rate. The last phase is phase four, evaluation of manual transcription and against automatic transcription and phonetic labelling using WER and FAR. The result of evaluation ASR engine using manual transcription serve as benchmark for acceptable accuracy ASR engine using automatic transcription and phonetic labelling. The acceptable accuracy of ASR engine using automatic transcription and phonetic labelling would be essential for development ASR system in future due to limitation using manual transcription. The detail result of this work from training and evaluations accuracy ASR engine are discussed in Chapter 4.

CHAPTER FOUR
ANALYSIS RESULTS

4.1 Introduction

This chapter discusses findings of ASR engine that have been trained for several times to gain optimum accuracy. However, training results depends on many factors such as vocabulary size, amount of data for training, type of words, and channel variability (Hosom, 2009). As has been mentioned in Chapter 3 accuracy of ASR engine were measured using two standard metrics WER and FAR. In this chapter the discussions emphasize in results of trainings from two ASR engines; ASR engine using manual transcription and ASR engine using automatic transcription and phonetic labelling.

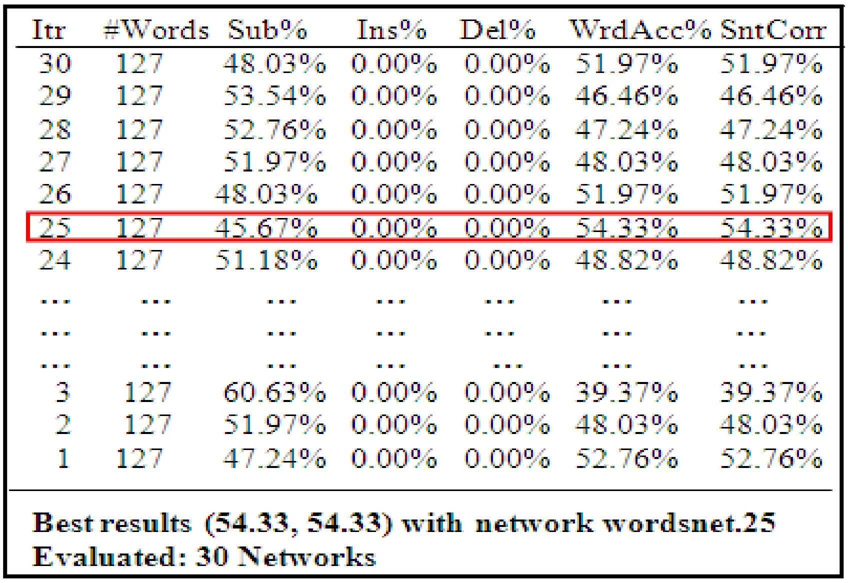
4.2 Trainings Result

Trainings are done to develop ASR engine for purpose of evaluation. Thus, investigating the accuracy of ASR engine is important because through the optimum accuracy of ASR engine can make sure any reading of dyslexic children’s enable to detect either it is right or wrong words.

The result of trainings was done after executing select_best.tcl with 30 iterations and once the recognizer is developed, the best network that gives the best result were chosen for the next training (second training). Since the training of recognition accuracy of ASR engine counting ‘errors’ substitution (i.e. Substituted of word “benang” for “menang”) so, substitution (sub %) also make impression for accuracy performances of ASR engine. The lower percentage of substitution (sub%) give the higher accuracy rate for ASR engine. Table 4.1 and Table 4.2 respectively depicted the first results of training ASR engine using manual transcription and ASR engine using automatic transcription and phonetic labelling. The WrdAcc% column shows word level accuracy, and the SntCorr% column depicted the percentage accuracy of “sentences”. Because of these training using only isolated words in BM not the sentences, so the result of SntCorr% also produced the same results performances with WrdAcc%.

The output for every training are in the form of a network file named wordsnet or wordsfanet1. In this study all trainings performed 30 iterations, the outputs of each of iterations is produced automatically named as wordsnet.1, wordsnet.2, wordsnet.3, wordsnet.4 until wordsnet.30. The same goes to wordsfanet1. The best accuracy using both technique transcriptions for first attempt training are highlighted in Table 4.1. and Table 4.2. The best accuracy for first training ASR engine using manual transcription is 54.33% given by `wordsnet.25.` Then, the best accuracy for first training using automatic transcription and phonetic labelling is 52.34% given by `wordsnet.12.` The result accuracy of automatic transcription and phonetic labelling given less than manual ones because of transcription and phonetic labelling files using force alignment is difficult to recognized at phoneme level and possibly result of accuracy performances automatic transcription and phonetic labelling slightly less than manual transcription.

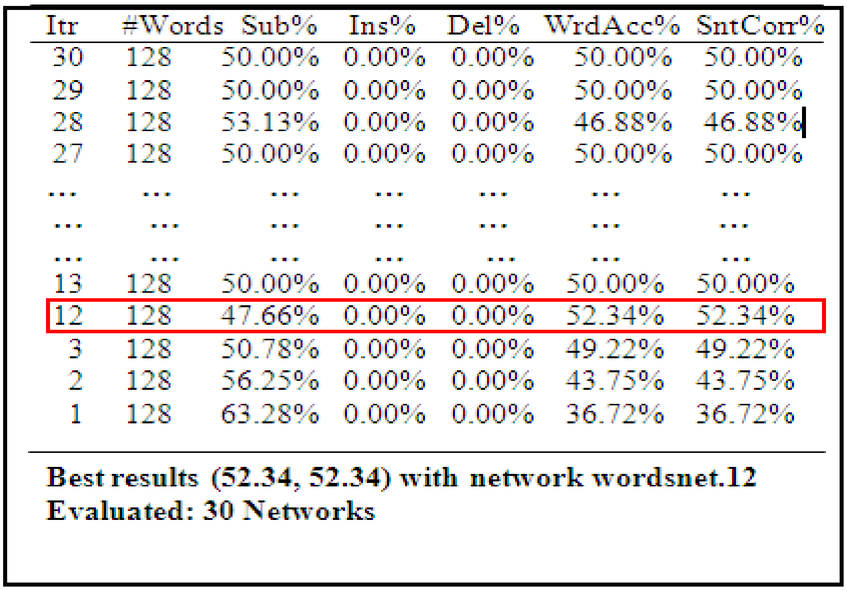
Table 4.1. First result training of ASR engine using manual transcription.



Itr	#Words	Sub%	Ins%	Del%	WrdAcc%	SntCorr
30	127	48.03%	0.00%	0.00%	51.97%	51.97%
29	127	53.54%	0.00%	0.00%	46.46%	46.46%
28	127	52.76%	0.00%	0.00%	47.24%	47.24%
27	127	51.97%	0.00%	0.00%	48.03%	48.03%
26	127	48.03%	0.00%	0.00%	51.97%	51.97%
25	127	45.67%	0.00%	0.00%	54.33%	54.33%
24	127	51.18%	0.00%	0.00%	48.82%	48.82%
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
3	127	60.63%	0.00%	0.00%	39.37%	39.37%
2	127	51.97%	0.00%	0.00%	48.03%	48.03%
1	127	47.24%	0.00%	0.00%	52.76%	52.76%

Best results (54.33, 54.33) with network wordsnet.25
Evaluated: 30 Networks

Table 4.2. First result training of ASR engine using automatic transcription and phonetic labelling.



Itr	#Words	Sub%	Ins%	Del%	WrdAcc%	SntCorr%
30	128	50.00%	0.00%	0.00%	50.00%	50.00%
29	128	50.00%	0.00%	0.00%	50.00%	50.00%
28	128	53.13%	0.00%	0.00%	46.88%	46.88%
27	128	50.00%	0.00%	0.00%	50.00%	50.00%
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
13	128	50.00%	0.00%	0.00%	50.00%	50.00%
12	128	47.66%	0.00%	0.00%	52.34%	52.34%
3	128	50.78%	0.00%	0.00%	49.22%	49.22%
2	128	56.25%	0.00%	0.00%	43.75%	43.75%
1	128	63.28%	0.00%	0.00%	36.72%	36.72%

Best results (52.34, 52.34) with network wordsnet.12
Evaluated: 30 Networks

Using the procedures of training in Chapter 3 the transcription and phonetic labelling for both technique transcription files must be re-trained to obtain optimum accuracy of ASR engine. This result would be re-trained to improve accuracy rate by using the same input files (.phn, .wav, and .txt), but different info file is used where a new partition name must be created and path category files must changed to “require: wt” where w is .wav files and t is .txt files. The important part also should be replace in this info file is “force_cat” because for the second training or next training the best network of previous training is used for possibly improve accuracy of ASR engine. Refer Figure 4.1 which some content of file have been modified for the next training.

```
basename:          words;
partition:         training2;
sampling_freq:     16000;
frame_size:        10;
min_sample:        100;

corpus name:       bmwords
cat_path:          bmwords_training2
require:           wt
partition:         "{expr $ID % 5} {0 1 2}"
filter:            1+1
lexicon:           words.lexicon
force_cat:         "fa.tcl wordsnet.25
                   words.training.spec words.lexicon
                   WAV TXT c OUT"
want:              ALL;
```

Figure 4.1. Info file is used for re-train ASR engine process.

4.3 Comparison Accuracy of ASR Engines using Manual and Automatic Transcription and phonetic labelling.

Using the best result of first training development dataset the experiments continue for second training through the same process using hybrid HMM/ANN. Then, the best network that has given the best result in the second training is used for next training and so on. The training would stop after the accuracy results of training development data set showed a decrease. The best results and best network have been chosen from several trainings of ASR engine using manual and automatic transcription and phonetic labelling that would be compared in terms of their accuracy performances. Refer Table 4.3 the findings of trainings using both transcription approach.

Table 4.3. The findings results of trainings using both transcription approach.

Training	**Manual transcription**		**Automatic transcription**	
	Best results WrdAcc%	**Best Network**	**Best results WrdAcc%**	**Best Network**
First	54.33%	Wordsnet.25	52.34%	Wordsnet.12
Second	58.27%	Wordsfa2net.30	56.25%	Wordsfa2net.4
Third	62.20%	Wordsfa3net.29	61.72%	Wordsfa3net.1
Fourth	61.42%	Wordsfa4net.18	59.38%	Wordsfa4net.24
Fifth	**76.29%**	**Wordsfa4net.29**	**76.04%**	**Wordsfa3net.9**

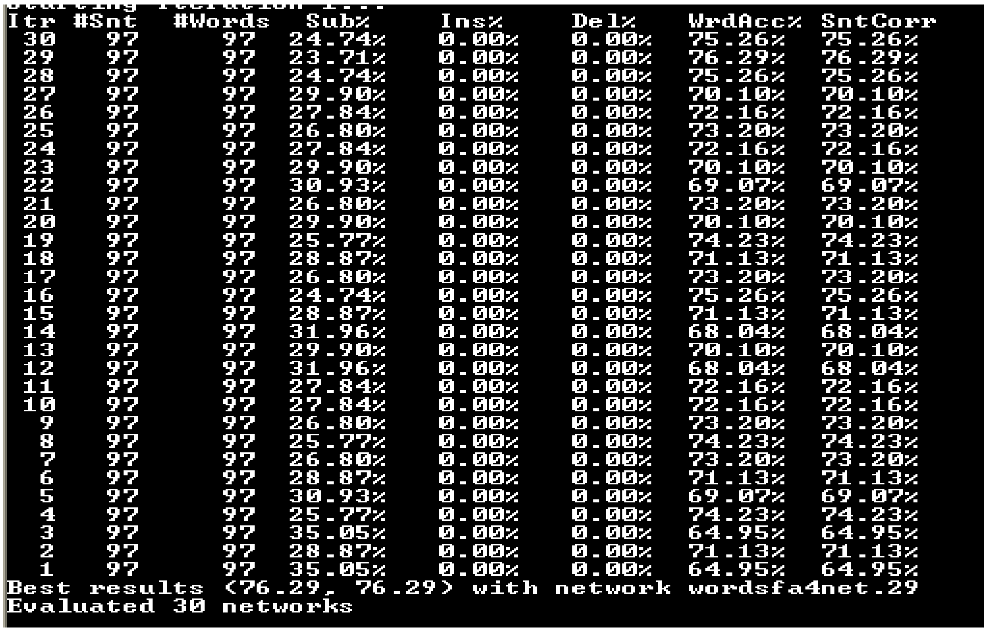
The trainings involved 30 networks iterations for both transcriptions files. The process iterates until optimum accuracy is achieved on the development dataset and only then it is tested on the dataset to evaluate final network. The ASR engine regarded the final network with highest recognition accuracy on test dataset that can be used for further evaluation using WER and FAR.

Referring on the results manual transcription training, the accuracy from the first until the third training shows improvements which are the first training is 54.33% has increased to 58.27%. Subsequently, the result of training using manual transcription increased 3.93% in the third training given 62.20%. The results are shows enhancement because of while training some adjustment is done on ‘tie’ category and prior perform each training a lexicon file is cleaned or modified to boost improvement accuracy rate. However, the performances recognition rate for fourth training was decreased to 61.42%. Thus, we need to stop the training on development dataset for manual transcription. The fifth training is final results for ASR engine using test dataset. The training in the test dataset used `Wordsfa3net.29` from the third training as input network to train final result for ASR engine using manual transcription. The result of ASR engine trained on manual transcription is 76.29%.

After that, the training was performed for ASR engine using automatic transcription and phonetic labelling. Based on Table 4.3 results of trainings ASR engine using automatic transcription and phonetic labelling is at par with results of manual transcription training. The first until third training results showed enhancement respectively 52.34%, 56.25% and 61.72%. The improvement in this training also using similar technique has been done by training using manual transcription files by changing tie 'category' and modified lexicon model every time training is performed. However, in the fourth training the results reduce to 59.38%. Therefore, the best results training development dataset for automatic transcription and phonetic labelling also in the third training. We used the third network `Wordsfa3net.1` as the input to train test data set because it given the best result. The accuracy of ASR engine for automatic transcription and phonetic labelling is 76.04% which is similar with ASR using manual transcription.

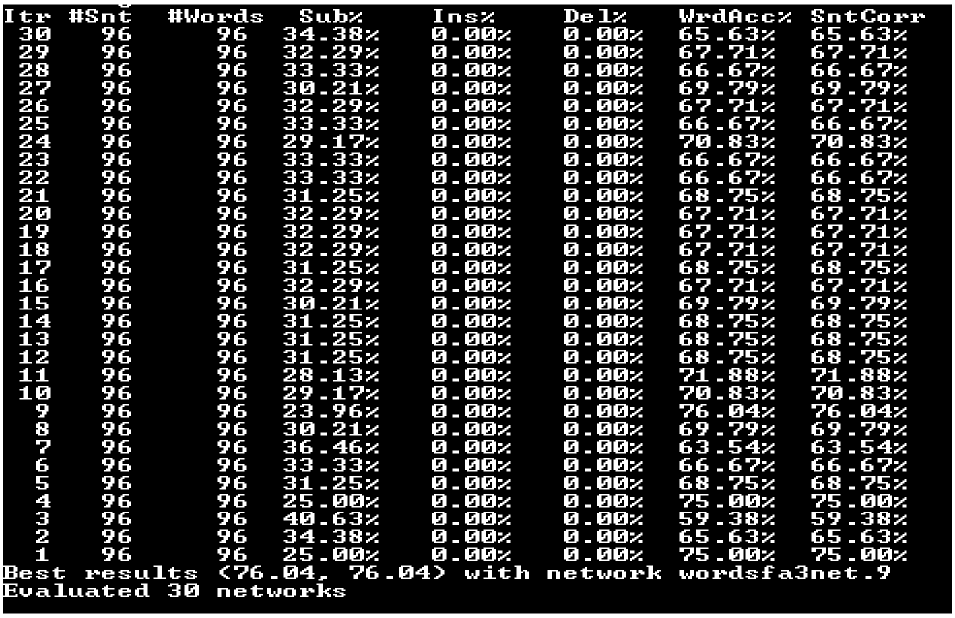
4.4 Evaluation WER and FAR

After trainings are done for manual and automatic transcription and phonetic labelling, then we proceed to evaluated WER and FAR using the test dataset training. For WER the training using test dataset as final result of two ASR engine are used. Figure 4.2 and Figure 4.3 depict performances of test dataset from best result in development data set to produce final result for evaluation of WER and FAR.



```
Itr #Snt  #Words   Sub%     Ins%     Del%    WrdAcc%  SntCorr
 30   97      97  24.74%   0.00%    0.00%   75.26%   75.26%
 29   97      97  23.71%   0.00%    0.00%   76.29%   76.29%
 28   97      97  24.74%   0.00%    0.00%   75.26%   75.26%
 27   97      97  29.90%   0.00%    0.00%   70.10%   70.10%
 26   97      97  27.84%   0.00%    0.00%   72.16%   72.16%
 25   97      97  26.80%   0.00%    0.00%   73.20%   73.20%
 24   97      97  27.84%   0.00%    0.00%   72.16%   72.16%
 23   97      97  29.90%   0.00%    0.00%   70.10%   70.10%
 22   97      97  30.93%   0.00%    0.00%   69.07%   69.07%
 21   97      97  26.80%   0.00%    0.00%   73.20%   73.20%
 20   97      97  29.90%   0.00%    0.00%   70.10%   70.10%
 19   97      97  25.77%   0.00%    0.00%   74.23%   74.23%
 18   97      97  28.87%   0.00%    0.00%   71.13%   71.13%
 17   97      97  26.80%   0.00%    0.00%   73.20%   73.20%
 16   97      97  24.74%   0.00%    0.00%   75.26%   75.26%
 15   97      97  28.87%   0.00%    0.00%   71.13%   71.13%
 14   97      97  31.96%   0.00%    0.00%   68.04%   68.04%
 13   97      97  29.90%   0.00%    0.00%   70.10%   70.10%
 12   97      97  31.96%   0.00%    0.00%   68.04%   68.04%
 11   97      97  27.84%   0.00%    0.00%   72.16%   72.16%
 10   97      97  27.84%   0.00%    0.00%   72.16%   72.16%
  9   97      97  26.80%   0.00%    0.00%   73.20%   73.20%
  8   97      97  25.77%   0.00%    0.00%   74.23%   74.23%
  7   97      97  26.80%   0.00%    0.00%   73.20%   73.20%
  6   97      97  28.87%   0.00%    0.00%   71.13%   71.13%
  5   97      97  30.93%   0.00%    0.00%   69.07%   69.07%
  4   97      97  25.77%   0.00%    0.00%   74.23%   74.23%
  3   97      97  35.05%   0.00%    0.00%   64.95%   64.95%
  2   97      97  28.87%   0.00%    0.00%   71.13%   71.13%
  1   97      97  35.05%   0.00%    0.00%   64.95%   64.95%
Best results (76.29, 76.29) with network wordsfa4net.29
Evaluated 30 networks
```

Figure 4.2. Test-dataset ASR engine using manual transcription.



```
Itr #Snt  #Words   Sub%     Ins%     Del%    WrdAcc%  SntCorr
 30   96      96  34.38%   0.00%    0.00%   65.63%   65.63%
 29   96      96  32.29%   0.00%    0.00%   67.71%   67.71%
 28   96      96  33.33%   0.00%    0.00%   66.67%   66.67%
 27   96      96  30.21%   0.00%    0.00%   69.79%   69.79%
 26   96      96  32.29%   0.00%    0.00%   67.71%   67.71%
 25   96      96  33.33%   0.00%    0.00%   66.67%   66.67%
 24   96      96  29.17%   0.00%    0.00%   70.83%   70.83%
 23   96      96  33.33%   0.00%    0.00%   66.67%   66.67%
 22   96      96  33.33%   0.00%    0.00%   66.67%   66.67%
 21   96      96  31.25%   0.00%    0.00%   68.75%   68.75%
 20   96      96  32.29%   0.00%    0.00%   67.71%   67.71%
 19   96      96  32.29%   0.00%    0.00%   67.71%   67.71%
 18   96      96  32.29%   0.00%    0.00%   67.71%   67.71%
 17   96      96  31.25%   0.00%    0.00%   68.75%   68.75%
 16   96      96  32.29%   0.00%    0.00%   67.71%   67.71%
 15   96      96  30.21%   0.00%    0.00%   69.79%   69.79%
 14   96      96  31.25%   0.00%    0.00%   68.75%   68.75%
 13   96      96  31.25%   0.00%    0.00%   68.75%   68.75%
 12   96      96  31.25%   0.00%    0.00%   68.75%   68.75%
 11   96      96  28.13%   0.00%    0.00%   71.88%   71.88%
 10   96      96  29.17%   0.00%    0.00%   70.83%   70.83%
  9   96      96  23.96%   0.00%    0.00%   76.04%   76.04%
  8   96      96  30.21%   0.00%    0.00%   69.79%   69.79%
  7   96      96  36.46%   0.00%    0.00%   63.54%   63.54%
  6   96      96  33.33%   0.00%    0.00%   66.67%   66.67%
  5   96      96  31.25%   0.00%    0.00%   68.75%   68.75%
  4   96      96  25.00%   0.00%    0.00%   75.00%   75.00%
  3   96      96  40.63%   0.00%    0.00%   59.38%   59.38%
  2   96      96  34.38%   0.00%    0.00%   65.63%   65.63%
  1   96      96  25.00%   0.00%    0.00%   75.00%   75.00%
Best results (76.04, 76.04) with network wordsfa3net.9
Evaluated 30 networks
```

Figure 4.3. Test dataset ASR engine using automatic transcription and phonetic labelling.

As we can see in Figure 4.2 and Figure 4.3 the final result of automatic transcription and phonetic labelling is 76.04% which is similar with manual transcription of 76.29%. The ASR engine using manual transcription, as always mentioned in the previous chapter, is the benchmark for accuracy of automatic transcription and phonetic labelling. Based on Figure 4.2 the finding result of ASR engine using automatic transcription and phonetic labelling is at par with ASR engine using manual transcription. So, automatic transcription and phonetic labelling can be used as alternatif to solve limitation problem using manual transcription.

Hence, the evaluation of ASR engine using manual transcription and ASR engine using automatic transcription and phonetic labelling are performed after training the test data set. The evaluation based on two metrics which are WER and FAR. The following is the way to calculate WER and FAR based on best result of word accuracy from the final result from both methods. Table 4.4 illustrates the calculation of WER and FAR for ASR engine using manual and ASR engine using automatic transcription and phonetic labelling. Figure 4.4 illustrates graph comparison between both methods.

Table 4.4. Calculation of WER and FAR for manual and automatic transcription and phonetic labelling.

Training ASR engines	**WER = 100% - Recognition rate**	**FAR = no. of correct reading recognized as incorrect / total no. of correct readings *100**
Manual	100% - 76.29% **= 23.7%**	23/ 128 *100 **=17.9%**
Automatic	100%-76.04 **= 23.9%**	23/128 *100 **=17.9%**

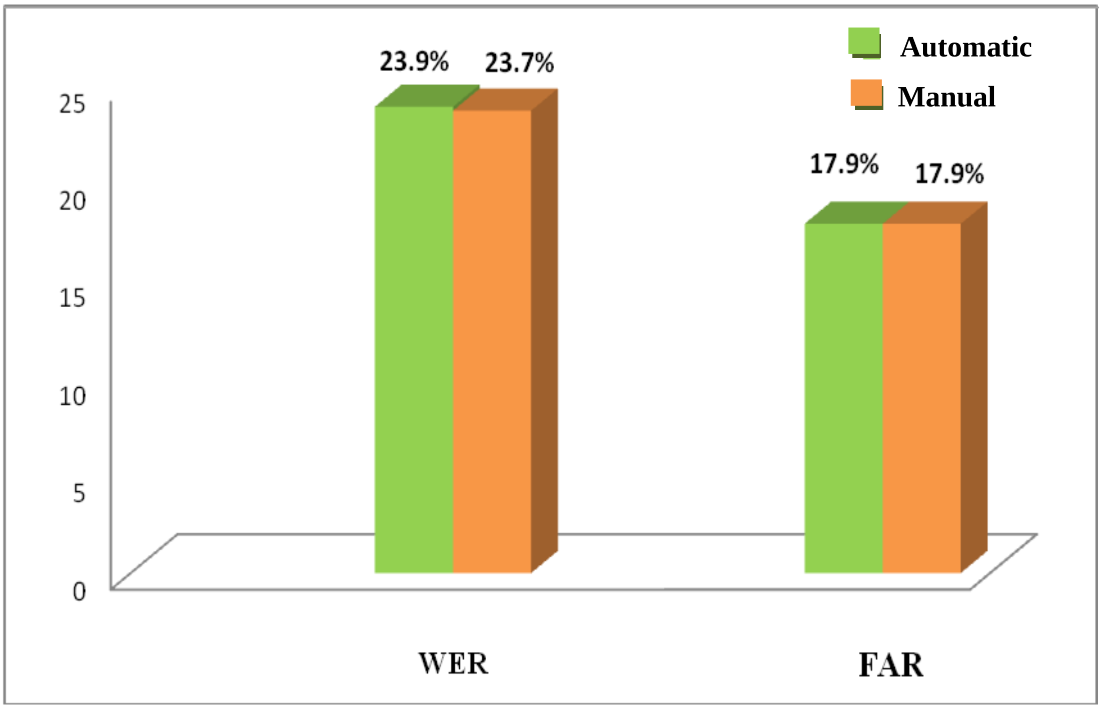


Figure 4.4. Graph comparison between both methods on evaluation WER and FAR.

Given the observation above, it is shown that since automatic transcription and phonetic labelling has similar WER with manual transcription, it can potentially be used to perform transcription for dyslexic children's read speech through forced alignment. In the WER evaluation, the recognition accuracy performances depend on recognizing the words (the lower percentages, the better accuracy). Therefore, the lowest WER for automatic transcription and phonetic labelling is 23.9% and manual transcription with 23.7%. The WER is influenced by highly phonetically similar errors from dyslexic children's speech that affected substitution (Sub). However, in this training the insertion error percentages (*Ins%*) and the deletion error percentage (*Del%)* is always zero as these two percentages are for sentence levels, so they are not calculated sentence level. Phonetic errors of dyslexic children's read speech is high thus affected the recognition accuracy of WER and FAR.

The FAR is also measured to support WER. Thus, to calculate FAR we need to divide number of readings recognized as incorrect with total number of correct readings as mentioned in Table 4.4. A fair comparison is seen in graph at figure 4.4 where the automatic transcription and phonetic labelling has misrecognized 23 correct reading as incorrect out of a total of 128 which is similar with manual transcription namely 17.9%. But, this recognizer performances result is still reliable even when dealing with highly phonetically similar errors words of dyslexic children speech.

4.5 Summary

Based on the result, we observed that the result obtained by ASR engine using automatic transcription and phonetic labelling training is 76.04%. This is similar to that performed by ASR engine using manual transcription, which is 76.29%. Thus, the WER for automatic transcription and phonetic labelling training is 23.9%, which has been developed using dyslexic children’s read speech with high phonetically similar errors in BM words. With that, this chapter has answered the questions, mentioned in Chapter 1- Can automatic transcription and phonetic labelling produce acceptable accuracy when dealing with highly phonetically similar errors of dyslexic’s children read speech in BM? In conclusion, the accuracy rate of automatic transcription and phonetic labelling can be used for children with dyslexia. Answering questions means that the final objective has been fulfilled that is to evaluate ASR engine using WER and FAR, as demonstrated. The final chapter concludes the research and discusses future work.

CHAPTER FIVE
CONCLUSION AND FUTURE WORK

5.1 Introduction

The study was set out to investigate accuracy of ASR engine using automatic transcription and phonetic labelling of dyslexic children's read speech in BM. The accuracy of ASR engine using automatic transcription and phonetic labelling has been evaluated whether it is acceptable for the purpose of development ASR engine for dyslexic children reading. Additionally, this study is significant to overcome limitation of manual transcription due to time consuming, costly, tedious and human transcribers tend to performed error while involved large speech files especially when dealing with highly phonetically similar errors. The highly phonetically similar errors of dyslexic children's read speech is a challenge to get optimum accuracy of ASR engine. Therefore, the study sought to answer the research question in investigating the accuracy of automatic transcription and phonetic labelling using forced alignment of dyslexic children's reading in BM. The question has been answered in this study where automatic transcription and phonetic labelling can produce acceptable accuracy when dealing with highly phonetically similar errors of dyslexic children read speech. The evidence of similarity results between ASR engine using automatic transcription and phonetic labelling and ASR engine using manual transcription prove that the automatic one is at par with the manual one.

5.2 Summary of the Thesis

In investigating whether or not automatic transcription and phonetic labelling can be used to development of ART and IRT for dyslexic children reading. It is important to realize, automatic transcription and phonetic labelling using forced alignment also facing problem to transcribe the word. This is because, the pattern of dyslexic children's read speech that contains highly phonetically similar errors giving difficulty for ASR to produce acceptable accuracy. The 585 speech files of dyslexic children's reading in BM performed transcription and phonetic labelling through two techniques; manual transcription by hand labelling and automatic transcription and phonetic labelling using forced alignment. The challenging involve in transcribing speech files using manual transcription are time consuming, tedious, and error prone due to highly phonetically similar errors that make us to be patient during manual transcription and phonetic labelling. Three months were taken to accomplish 585 speech files using manual transcription.

Thus, both transcription files that are using different techniques need to go through training process using the state of the art techniques, HMM-ANN. The training process of transcription and phonetic labelling files constructed the ASR engine at the same time produced accuracy of recognition rate of the ASR engine. In this study, the accuracy between ASR engine using manual transcription and ASR engine using automatic transcription and phonetic labelling are compared. Based on accuracy of ASR engine using manual transcription as reference, acceptable accuracy of ASR engine using automatic transcription and phonetic labelling is measured.

Surprisingly, the accuracy of both automatic transcription and phonetic labelling and manual transcription is similar, given 76.04% and 76.26% respectively. The previous study by Hosom (2002) obtained the results of automatic transcription is 97.24% much similar with manual transcription which are 97.54%. Although, Hosom (2002) obtained high accuracy performance of both transcription but we realize he is using normal person without dealing with dyslexic children's highly phonetically similar errors. These results in this study are still acceptable because according to Kheir and Way (2006) accuracy of reasonably well-trained ASR systems typically is around 75% upwards.

The recognition rate from both ASR engine has been evaluated using WER and FAR. In the WER evaluation, the lower percentage is the better. Therefore, the WER of ASR engine using automatic transcription and phonetic labelling is 23.9%. Then, for ASR engine using manual transcription is 23.7%. The FAR were support WER performances and to measure FAR we divide number of recognized as in correct reading with the total number of correct. The result FAR for both transcriptions and phonetic labelling are similar is 17.9%.

5.3 Contribution of the Study

The result obtained of ASR engine using automatic transcription and phonetic labelling look similar with results of ASR engine using manual transcription given significance for us to solving the issues transcribing and labelling speech files using human transcribers. The lexicon model of this study also can be used to further expand ASR engine using automatic transcription and phonetic labelling. The best results of the best network from the training process to construct ASR engine can be used to transcribe and labelling phonetic symbols for larger speech files using forced alignment.

This dissertation has proven that automatic transcription and phonetic labelling can still give acceptable accuracy given the nature of dyslexic children's read speech. Using automatic transcription and phonetic labelling given acceptable accuracy performance would help researchers to develop ASR engine using automatic transcription and phonetic labelling. Moreover, the results are obtained in this study would help children with dyslexia using ASR system in BM. This is because study area in development ASR system in BM is still in its infancy in Malaysia having the automatic approach would facilitate. The development of ASR system in BM, which can give benefit to Malaysian children who has learning disabilities (e.g. dyslexic children and children with literacy problem).

5.4 Future Work

In correspondence for improvement accuracy of ASR engine, the future work could be carried out to expand the vocabulary that consist more words in BM. Using larger speech files or corpus act as input for ASR system would help transcription and phonetic using forced alignment to finding the similar feature of speech children with dyslexia in terms of similar signal (waveform), phonetic symbols and all boundaries related to words that ASR tried to be matching. Moreover, by adding vocabulary in BM for transcription and phonetic labelling hopefully can construct ASR engine that obtain higher recognition rate for dyslexic children. The high accuracy of ASR engine in BM also can be used for general purpose like developing ART and IRT for Malaysian children, telephone recognition, and source for phonetic research.

5.5 Concluding Remarks

This dissertation presents the automatic transcription and phonetic labelling using forced alignment that take dyslexic children's speech reading in BM as input to perform transcription and phonetic labelling. The study also has achieved all the research objectives stated in Chapter One: the first objective to produce manual transcription and phonetic labelling; secondly objective to construct automatic transcription and phonetic labelling using forced alignment; and the third objective is to evaluated automatic transcription and phonetic labelling using forced alignment against manual transcription using WER and FAR. These objectives answered the research question that demonstrate that ASR engine using automatic transcription and phonetic labelling produce acceptable accuracy even though dealing with highly phonetically similar errors. The technique for manual and automatic transcription and phonetic labelling have been discussed including issues using manual transcription, HMM-ANN for development of ASR engine, and standard metrics WER and FAR for evaluation accuracy performances. Transcription and phonetic labelling is the basic element for developing an ASR engine. So, potential automatic transcription and phonetic labelling using forced alignment are faster, saving time and lower cost to transcribing speech files make it become important for development of any ASR system in BM. As conclusion, although ASR recognition system in BM is still in its infancy but the significant of this study would help enhance ASR system in BM to be gradually developed that can have impact on children with dyslexia to use it to help their learning process.

REFERENCES

Abushariah, A. A. M., Gunawan, T. S., Khalifa, O. O., Abushariah, M. A. M. (2010). English digits speech recognition system based on Hidden Markov Models. In *International Conference on Computer and Communication Engineering (ICCCE)*, Kuala Lumpur, Malaysia.

Al-Manie, M. A., Alkanhal, M. I., & Al-Ghamdi, M. M. (2009). Automatic speech segmentation using the Arabic phonetic database. In *Proceedings of the World Scientific and Engineering Academy and Society (WSEAS), Automation & Information*, *10*, 6-79.

Athanaselis, T., Bakamidis, S., Dologlou, I., Argyriou, E. N., & Symvonis, A. (2014). Making assistive reading tools user friendly: a new platform for Greek dyslexic students empower by automatic speech recognition. *Multimedia Tools and Application*, *68*(3), 681-699.

Azam, S. M., Mansoor, Z. A., Mughal, M. S., & Mohsin, S. (2007). Urdu spoken digits recognition using classified MFCC and backprogation neural network. In *Computer Graphics, Imaging and Visualisation, IEEE, 7*, 414-418.

Banerjee, S., Beck, J. E., & Mostow, J. (2003). Evaluating the Effect of Predicting Oral Reading Miscues. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), 8*.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2000). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1), 5-22.

Bauer, T., Hitzenberger, L., & Hennecle, L. (2002). Effects of manual phonetic transcriptions on recognition accuracy of streetnames. In *Proceedings of the International Symposiums for Information Swissenschaft (ISI)*, *8*, 21-25.

Bhotto, M. Z. A., & Amin, M. R. (2004). Bengali text dependent speaker identification using melfrequency cepstrum coefficient and vector quantization. In *International Conference on Electrical & Computer Engineering (ICECE)*, *3*, 28-30.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (version 5.4.08) [computer program]. Retrieved April 11, 2015, from http://www.fon.hum.uva.nl/praat/manual/Intro.html.

Bourassa, D., & Treiman, R. (2003). Spelling in children with Dyslexia: Analysis from the Treiman-Bourassa Early spelling test. *Scientific studies of reading, 7*(4), 309-333.

Bourlard, H. A., & Morgan, N. (2012). Connectionist speech recognition: A hybrid approach. *Springer Science & Business Media*, *247.*

Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. (2012). Train & Align: A new online tool for automatic phonetic alignment. In *IEEE Workshop on Spoken Language Technologies*, 416-421.

Cangemi, F., Cutugno, F., Ludusan, B., Seppi, D., & Van C. D. (2011). Automatic Speech Segmentation for Italian (Assi): Tools, Models, Evaluation, and Applications. In *Proceedings of the Associazione Italiana di Scienze della Voce (AISV)*, Lecce, Italy, *7*, 337-344.

Carroll, J. M., & Myers, J. M. (2010). Speech and language difficulties in children with and without a family history of dyslexia. *Scientific Studies of Reading*, *14*(3), 247-265.

Castles, A., Wilson, K., & Coltheart, M. (2011). Early orthographic influences on phonemic awareness tasks: evidence from a preschool training study. *Journal of Experimental Child Psychology, 108*(1), 203-210.

Chang, S., Shastri, L., & Greenberg, S. (2000). Automatic Phonetic transcription of spontaneous speech (American English). In *Proceedings of the International conferences on Spoken Languages Processing,* Beijing, China, *6*, 330-333.

Chou, F. C., Tseng, C. Y., & Lee, L. S. (2002). A set of corpus-based text-to-speech synthesis technologies for Mandarin Chinese. *Speech and Audio Processing, IEEE Transactions on, 10*(7), 481-494.

Conn, N., & McTear, M. (2000). Speech Technology: A Solution for People with Disabilities. In *IEEE Seminar on Speech and Language Processing for Disabled and Elderly People, 7,* 1-6.

Cosi, P., & Hosom, J. P. (1999). Hmm/Neural Network-Based System for Italian Continuous Digit Recognition. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS), 14,* 1669-1672.

Choudhary, A., Chauhan, M. R., & Gupta, M. G. (2013). Automatic speech recognition system for isolated & connected words of Hindi language by using Hidden Markov Model Toolkit (HTK). In *Proceedings of the International Conference on Emerging Trends in Engineering and Technology (ACEEE),* 847-853.

Cucchiarini.C., & Strik, H. (2003). Automatic phonetic transcription: An overview. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, Barcelona, *15*, 347–350.

Das, R., Izak, J., Yuan, J., & Liberman, M. (2010). Forced alignment under adverse conditions. *University of Pennsylvania, CIS Dept. Senior Design Project Report.*

DeFries, J. C., Olson, R. K., Pennington, B. F., & Smith, S. D. (1991). Colorado Reading Project: Past, present, and future. *Learning Disabilities: A Multidisciplinary Journal, 2,* 37-46.

Demuynck, K., & Laureys, T. (2002). A comparison of different approaches to automatic speech segmentation. In *Text, Speech and Dialogue, 5*, 277-284.

Dinarelli, M., Moschitti, A., & Riccardi, G. (2009). Concept Segmentation and Labeling for Conversational Speech. In *Annual Conference of the International Speech Communication Association*, *10*, 2747-2750.

Douklias, S., Masterson, J., & Hanley, J. R. (2010). Surface and phonological developmental dyslexia in Greek. *Cognitive Neuropsychology*, *26,* 705-723.

Dupuis, A. (2011). Automatic transcription of audio files and why manual transcription may be better. Retrieved March 23, 2015, from: http://www.researchware.com/company/blog/368-automatic transcription.html.

Evermann, G. (1999). Minimum word error rate decoding. *Cambridge University, UK*, 45-67.

Fadhilah, R., & Ainon, R., N. (2008). Isolated Malay speech recognition using Hidden Markov models. *Proceedings of the International Conferences on Computer and Communication Engineering*, 721-725.

Fang, C. (2009). From Dynamic Time Warping (DTW) to Hidden Markov Model (HMM). Final Project report, University of Cincinnati.

Fish, R., Hu, Q., & Boykin, S. (2006). Using audio quality to predict word error rate in an automatic speech recognition system. *Unpublished from MITRE corporation.*

Frikha, M., & Hamida, A. B. (2012). A comparative survey of ANN and hybrid HMM/ANN architectures for robust speech recognition. *American Journal of Intelligent Systems*, *2*(1), 1-8.

Gemello, R., Mana, F., & Albesano, D. (2010). Hybrid HMM/Neural Network based Speech Recognition in Loquendo ASR. Retrieved December, 2, 2014, from http://www. loquendo. com/en/.

Gianna, A., Mclaughlin, T. F., Derby K. M., & Waco, T. (2012). The effects of the Davis symbol mastery system to assist a fourth grader with dyslexia. In *Spelling: A Case Report. I-manager's Journal on Educational Psychology, 6*(2) 13-18.

Gibbon, D. (1997). Part 1: Spoken language system and corpus design. In *Handbook of standards and resources for spoken language systems*. Berlin: Mouton de Gruyter, 152.

Giurgiu, M., & Kabir, A. (2012). Automatic transcription and speech recognition of Romanian corpus RO-GRID. In *International Conference of the Telecommunications and Signal Processing (TSP), 35*, 465-468.

Goldman, J. P., & Schwab, S. (2014). Easyalign Spanish: An (Semi-) Automatic Segmentation Tool Under Praat. In *Salvador Plans, A. Fonética Experimental, Education Superior Investigation.* Madrid, *1*, 629-640.

Goldman, J. P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *Annual Conference of the International Speech Communication Association*, Folorence, *12*, 3233-3236.

Handler, S. M., & Fierson, W. M. (2011). Learning disabilities, dyslexia, and vision. *Paediatrics*, *127*(3), 818-856.

Hagen, A., Pellom, B., & Cole, R. (2003). Children's speech recognition with application to interactive books and tutors. In *Proceedings of the Automatic Speech Recognition and Understanding* (*ASRU*), *3*, 186-191.

Hagen, A. (2006). Advances in children’s speech recognition with application to interactive literacy tutors. Doctoral dissertation, University of Colorado.

Haykin, S. (1999). Neural networks: a comprehensive foundation. (2nd ed.) Upper Saddle Rever, New Jersey: Prentice Hall.

Hazen, T. J. (2006). Automatic alignment and error correction of human generated transcripts for long speech recordings. *Proceedings of International Conference on Spoken Language Processing,* Pittsburgh, *9*, 1606-1609.

Hieronymus, L. J. (1993). ASCII Phonetic Symbols for the world's Languages: Worldbet, Bell laboratories manuscript.

Hofmann, S., & Pfister, B. (2010). Fully automatic segmentation for prosodic speech corpora. In *Eleventh Annual Conference of the International Speech Communication Association*, Makuhari, Japan, 1389-1392.

Hosom, J. P. (2002). A Comparison of speech recognizers created using manually-aligned and automatically-aligned training data. *Technical Report CSE-00-02,* Oregon Graduate Institute of Science and Technology, Center for spoken Language Understanding, Beaverton.

Hosom, J. P. Shriberg, L., & Green, J. R. (2004). Diagnostic assessment of childhood apraksia of speech using automatic speech recognition (ASR) methods. *Journal of medical speech-language pathology*, *12*(4), 167.

Hosom, O., Villiers, J., Cole, R., Fanty, M., Schalkwyk, J., Yan, Y., & Wei, W. (2006). Training HMM/ANN Hybrids for Automatic Speech Recognition. Retrieved July 3, 2014, from http://www.cslu.ogi.edu/tutordemos/nnet_training/tutorial.html

Hosom, J. P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, *51*(4), 352-368.

Husniza, H., & Zulikha, J. (2009). Dyslexic children's reading pattern as input for ASR: Data, analysis, and pronunciation model. *Journal of Information and Communication Technology*, *8*, 1-13.

Husniza, H. (2010). Automatic speech recognition model for dyslexic children reading in bahasa Melayu. Doctoral dissertation, Universiti Utara Malaysia.

Husniza, H., & Zulikha, J. (2010). Improving ASR performances using context-dependent phoneme models. *Journal of Systems and Information Technology (JSIT)*, *12*(1), 56-69.

Husniza, H., Yuhanis, Y., & Siti Sakira, K. (2013a). Speech Malay language influence on automatic transcription and segmentation. *Proceeding of the International Conferences on Computing and Informatics*, ICOCI, Sarawak, Malaysia, *4*, 132-137.

Husniza, H., Yuhanis, Y., & Siti Sakira, K. (2013b). Evaluation of phonetic labeling and segmentation for dyslexic children's speech. *Proceeding of the World Congress one Engineering*, London, U.K, *2*.

Jackson, M. (2005). Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language. Master dissertation, Computer Science of Makerere University.

Jakovljevic, N., Miskovic, D., Pekar, D., Secujski, M., & Delic, V. (2012). Automatic Phonetic Segmentation for a Speech Corpus of Hebrew, *Infotch-Jahorina*, *11*, 742-745.

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech communication*, *45*(4), 455-470.

Jiang, F., Yuan, J., Tsaftaris, S. A., & Katsaggelos, A. K. (2011). Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, *115*(3), 323-333.

Jurafsky, D., & James, H. (2000). Speech and language processing: *An introduction to natural language processing, computational linguistics, and speech*. Prentice Hall, New Jersey, USA, *2*.

Kabir, A., Barker, J., & Giurgiu, M. (2010). Integrating hidden Markov model and PRAAT: a toolbox for robust automatic speech transcription. In *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments, 7745.*

Kaur, E. A., & Singh, E. T. (2010). Segmentation of continuous Punjabi speech signal into syllables. In *Proceedings of the World Congress on Engineering and Computer Science, 1*, 20-22.

Kawachale, M. S., & Chitode, J. S. (2012). Relative functional comparison of neural and non-neural approaches for syllable segmentation in Devnagari TTS system. *Proceedings of the International Journal of Computer Science Issues (IJCSI), 9*(3), 534-543.

Kawai, H., & Toda, T. (2004). An evaluation of automatic phone segmentation for concatenative speech synthesis. In *Proceedings of the International Conference Acoustics, Speech, and Signal Processing (ICASSP'04), 1*, 677-680.

Kheir, R., & Way, T. (2006). Improving speech recognition to assist real time classroom note taking. In *Proceedings of Rehabilitation Engineering Society of North America (RESNA) Conference*, *29*, 1-4.

Kim, Y. J., & Gibbon, D. C. (2011). Automatic Learning in Content Indexing Service Using Phonetic Alignment. In *Annual Conference of the International Speech Communication Association*, *12*, 925-928.

Kimball, O., Kao, C. L., Arvizo, T., Makhoul, J., & Iyer, R. (2004). Quick transcription and automatic segmentation of the Fisher conversational telephone speech corpus. In *Proceedings of Rich Transcription Workshop*, Palisades, Newyork.

Kuo, J. W., & Wang, H. M. (2006). A minimum boundary error framework for automatic phonetic segmentation. In *Proceedings of the International Conference on Chinese Spoken Language Processing*. Springer-Verlag, *5*, 399-409.

Kuo, J. W., Lo, H. Y., & Wang, H. M. (2007). Improved HMM/SVM methods for automatic phoneme segmentation. In *Annual Conference of the International Speech Communication Association*, *8*, 2057-2060.

Kvale, K.(1993). Segmentation and Labeling of Speech. (A Dissertation The Doctoral Degree, *The Norwegian Institute of Technology*).

Lakra, S., Prasad, T. V., Sharma, D. K., Atrey, S. H., & Sharma, A. K. (2012). Application of fuzzy mathematics to speech-to-text conversion by elimination of paralinguistic content. In *Proceedings of National Conferences on Soft Computing and Artificial Intelligence*, *arXiv preprint arXiv:1209.4535*, 294-299.

Lee, C. C., Katsamanis, A., Black, M. P., Baucom, B. R., Georgiou, P. G., & Narayanan, S. S. (2011). Affective state recognition in married couples' interactions using PCA-based vocal entrainment measures with multiple instance learning. In *Proceedings of the International Conferences on Affective Computer Intelligent Interaction (ACII)*, *2*, 31-41.

Lee, K., Hagen, A., Romanyshyn, N., Martin, S., & Pellom, B. (2004). Analysis and detection of reading miscues for interactive literacy tutors. In *Proceedings of the international conference on Computational Linguistics*. Association for Computational Linguistics. *20*, 1254.

Lee, L. W. (2008). Development and validation of a reading-related assessment battery in Malay for the purpose of dyslexia assessment. *Annals of Dyslexia, 58*(1), 37-57.

Leither, C. (2008). Data-Based Automatic Phonetic Transcription. Diploma Thesis, Signal Processing and Speech Communication Lab Graz University of Technology.

Levy, C., Linares, G., Bonastre, J. F., Stepmind, S. A., & Cannet, L. (2005). Mobile phone embedded digit-recognition. In *Workshop on DSP in Mobile and Vehicular Systems,* Sesimbra, Portugal.

Li, X., Ju, Y. C., Deng, L., & Acero, A. (2007). Efficient and robust language modeling in an automatic children's reading tutor system. In *International Conference* on *Acoustics, Speech and Signal Processing (ICASSP), 4,* 193-196.

Li, X., Deng, L., Ju, Y. C., & Acero, A. (2008). Automatic children's reading tutor on hand-held devices. In *Annual Conference of the International Speech Communication Association*, *9*, 1733-1736.

Lin, C. Y., Jang, J. S. R., & Chen, K. T. (2005). Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS. *Computational Linguistics and Chinese Language Processing*, *10*(2), 145-166.

Lu, L., Ghoshal, A., & Renals, S. (2013). Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 374-379.

Mandal, S., Das, B., Mitra, P., & Basu, A. (2011). Developing Bengali speech corpus for phone recognizer using optimum text selection technique. *International Conference in Asian Language Processing (IALP)*, IEEE Computer Society. 268-271.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. In *Computational linguistics*, *19*(2), 313-330.

Martens, J. P., Binnenpoorte, D., Demuynck, K., Van P. R., Laureys, T., Goedertier, W., et al. (2002). Word Segmentation in the Spoken Dutch Corpus. In *International conference on Language Resources and Evaluation (LREC), 3,* 1432-1437.

McIntyre, C. W., & Pickering, J. P. eds. (1995). Clinical studies of multisensory structured language education. Dallas, TX: *International Multisensory Structured Language Education Council.*

Milde, B. (2014). Unsupervised acquisition of acoustic models for speech-to-text alignment. Master’s Thesis, University Technical Darmstat.

Mishra, T., Ljolje, A., & Gilbert, M. (2011). Predicting Human Perceived Accuracy of ASR Systems. In *Annual Conference of the International Speech Communication Association*, *12*, 1945-1948.

Mohammad, W., Ruzanna, W. M., Vijayaletchumy, S., Aziz, A., Yasran, A., & Rahim, N. A. (2011). Dyslexia in the aspect of Malay language spelling. *International Journal of Humanities and Social Science (IJHSS), 21*(1), 266-268.

Mostow, J. (2006). Is ASR accurate enough for automated reading tutors, and how can we tell? In *International Conference on Spoken Language Processing. (ICSLP), 9.*

Mporas, I., T. Ganchev, & Fakotakis, N. (2010). Speech segmentation using regression fusion of boundary predictions. *Computer Speech & Language*, *24*(2), 273-288.

Mustafa, M. B., Rosdi, F., Salim, S. S., & Mughal, M. U. (2015). Exploring the Influence of General and Specific Factors on the Recognition Accuracy of an ASR System for Dysarthric Speaker. *Expert Systems with Applications*, *42*, 3924-3932.

Naghibi, T., Hofmann, S., & Pfister, B. (2013). An efficient method to estimate pronunciation from multiple utterances. In *Interspeech Annual Conference of the International Speech Communication Association*, *14*, 1951-1955.

Necibi, K., & Bahi, H. (2012). An Arabic mispronunciation detection system by means of automatic speech recognition technology. In *the International Arab Conference on Information Technology Proceedings*, *13*, 304-308.

Newton, J. M., & Thomas, E. M. (1974). Dyslexia A Guide for Teachers and Parents. *London: University Press.*

Novotney, S., & Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language*

Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 207-215.

Ong, H. F., & Ahmad, A. M. (2011). Malay Language Speech Recognizer with Hybrid Hidden Markov Model and Artificial Neural Network (HMM/ANN). In *International Journal of Information and Education Technology*, *1*(2), 114-119.

Passy, C. (2008). Turning audio into words on the screen. Retrieved January 25, 2015, from http://www.wsj.com/articles/SB122351860225518093.

Pedersen, J. S., & Larsen, L. B. (2010). A Speech Corpus for Dyslexic Reading Training. *Proceedings of the International Conference on Language Resources and Evaluation* (*LREC),* European Language Resources Association, *7,* 2820-2823.

Perea, M., Jimenez, M., Suarez C. P., Fernandez, N., Vina, C., & Cuetos, F. (2014). Ability for voice recognition is a marker for dyslexia in children.

Picone, J., Ganapathiraju, A., & Hamaker, J. (2006). Applications of Kernel Theory to speech. Recognition. *Kernel Methods in Bioengineering, Signal and Image Processing*, 224-240.

Pieraccini, R. (2012). The voice in the machine: Building computers that understand speech Massachusetts Institute of Technology (MIT Press), Cambridge, *141*.

Rabiner, L. R., & Juang, B. H. (1993). Fundamentals of speech recognition, prentice-hall, Englewood.

Radi. M. I. H. (2012). Phonetic transcription: A comparison between manual and automated approach. Master Thesis's, Universiti Utara Malaysia.

Rahman, F. D., Mohamed, N., Mustafa, M. B., & Salim, S. S. (2014). Automatic speech recognition system for Malay speaking children. In *ICT International Student Project Conference (ICT-ISPC), 3*, 79-82.

Ramesh, K. V., & Gahankari, S. (2013). Hybrid Artificial Neural Network and Hidden Markov Model (ANN/HMM) for speech and speaker recognition. In *International conference on Green Computing and Technology*, 24-27.

Rapp, S. (1995). Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An Aligner for German. In *Proceedings of ELSNET Goest East and IMACS Workshop,* Moscow, Russia. Retrieved January, 23, 2015, from http://www.ims.uni-stuttgart.˜de/rapp/.

Rasmussen, M. H., Tan, Z. H., Lindberg, B., & Jensen, S. H. (2009). A System for Detecting Miscues in Dyslexic Read Speech. In *Annual Conference of the International Speech Communication Association*, *10,* 1467-1470.

Rello, L., & Llisterri, J. (2012). There are phonetic patterns in vowel substitution errors in texts written by persons with dyslexia. In *Annual World Congress on Learning Disabilities. Learning disabilities: Present and future,* Oviedo, Spain. *21*, 327-38.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., et al. (1999). Stochastic pronunciation modeling from hand-labelled phonetic corpora. *Speech Communication*, *29*(2), 209-224.

Rosdi, F., & Ainon, R. N. (2008). Isolated Malay speech recognition using Hidden Markov Models. *Proceedings of the International Conference on Computer and Communication Engineering*, 721-725.

Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bonham, B., et al. (1996). Application of automatic speech recognition to speech and language development in young children. In *Proceedings spoken language of the International Conference on Spoken Language Processing*, Philadelphia, *1*, 176-179.

Saraclar, M., & Khundanpur, S. (2004). Pronunciation change in conversational speech and its implications for automatic speech recognition. In *Computer, Speech and Language*, *18*, 375-395.

Sarma, H., Saharia, N., & Sharma, U. (2014). Development of Assamese speech corpus and automatic transcription using HTK. In *Advances in Signal Processing and Intelligent Recognition Systems*. Springer International Publishing, *264*, 119-132.

Sawyer, D. J., Wade, S., & Kim, J. K. (1999). Spelling errors as a window on variations in phonological deficits among students with dyslexia. *Annals of Dyslexia*, *49*, 137 - 159.

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*(1), 96-109.

Serridge, B. (2014). An Undergraduate Course on Speech Recognition Based on the CSLU Toolkit. *In International Conference on Spoken Language Processing,* Sydney, Australia, *5*.

Shire, M. L. (2001). Relating frame accuracy with word error in hybrid ANN-HMM ASR. In *Proceedings of the European Conference on Speech Communication and Technology*, *7,* 1797-1800.

Shrawankar, U., & Mahajan, A. (2013). Speech: A Challenge to Digital Signal Processing Technology for Human-to-Computer Interaction. *arXiv preprint arXiv:1305.1925*. 206-212.

Silber, V., & Geri, N. (2014). Can automatic speech recognition be satisfying for audio/video search? Keyword-focused analysis of Hebrew automatic and manual transcription. *Online Journal of Applied Knowledge Management*, *2*(1), 104-121.

Sjolander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik,* 93-96.

Sjolander, K., & Beskow, J. (2006). WaveSurfer user manual. Retrieved April 9, 2015, from https://www.speech.kth.se/wavesurfer/man.html.

Sperber, M. (2012). Efficient speech transcription through respeaking. Master's Thesis, Karlsruhe Institute of Technology Department of Computer Science.

Stolcke, A., Ryant, N., Mitra, V., Yuan, J., Wang, W., & Liberman, M. (2014). Highly accurate phonetic segmentation using boundary correction models and system fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 14*, 5552-5556.

Sutton, S., Cole, R. A., De Villiers, J., Schalkwyk, J., Vermeulen, P. J., Macon, M. W., et al. (1998). In *Proceedings of the International Conference on Spoken Language Processing (ICSLP), 98*, 3221-3224.

Taileb, M., Al-Saggaf, R., Al-Ghamdi, A., Al-Zebaidi, M., & Al-Sahafi, S. (2013). YUSR: speech recognition software for dyslexics. *Design, User Experience, and Usability. Health, Learning, Playing, Cultural, and Cross-Cultural User Experience,* Springer Berlin Heidelberg. *8013*, 296-303.

Ting, C. M. (2007). Malay continuous speech recognition using continuous density Hidden Markov Model. Doctoral dissertation, Faculty of Electrical Engineering, Universiti Teknologi Malaysia.

Ting, C. M., & Hussain, S. H., Tan, S. T., & Ariff, A. K. (2007). Automatic phonetic segmentation of Malay speech database. In *International Conference on Information, Communications & Signal Processing, 6*, 1-4.

Tjalve, M., & Huckvale, M. (2005). Pronunciation variation modelling using accent features. In *Proceedings of Euro Speech, Speech Communication, 50*, 605-615.

Togneri, R., Alder, M. D., & Attikiouzel, Y. (1990). Speech processing using artificial neural networks. In *Proceedings of the Australian International Conferences on Speech Science and Technology, 3,* 304-309.

Tolba, M. F., Nazmy, T., Abdelhamid, A. A., & Gadallah, M. E. (2005). A novel method for Arabic consonant/vowel segmentation using wavelet transform.

International Journal on Intelligent Cooperative Information Systems, IJICIS, *5*(1), 353-364.

Toth, L., & Kocsor, A. (2007). A segment-based interpretation of HMM/ANN hybrids. *Computer Speech and Language*, *21*, 562-578.

Van Bael, C., Boves, L., Heuvel, H. & Strik, H. (2007). Automatic Phonetic Transcription of Large Speech Corpora. *Centre for Language and Speech Technology (CLST)*, Netherlands, *21*(4), 652-668.

Vasilescu, I., Vieru, B., & Lamel, L. (2014). Exploring pronunciation variants for Romanian speech-to-text transcription. In *Spoken Language Technologies for Under-Resourced Languages (SLTU).*St. Petersburg, Russia, 162-168.

Vijayalakshmi, A. (2012). Implementation of Forced Alignment Algorithm For Large Malay Database. Undergraduate Project's Paper, Universiti Teknologi Malaysia.

Wang, Y. Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy? In *Automatic Speech Recognition and Understanding (ASRU)*. IEEE Workshop, *3*, 577-582.

Wells, J. C. (2006). Phonetic transcription and analysis. *Encyclopaedia of Language and Linguistics.* Amsterdam: Elsevier, 386-396.

Wester, M. (2003). Pronunciation modelling for ASR knowledge based and data derived methods. In *Computer Speech and Language, 17*(1), 69-85.

Williams, J. D., Melamed, I. D., Alonso, T., Hollister, B., & Wilpon, J. (2011). Crowd-sourcing for difficult transcription of speech. In *Automatic Speech Recognition and Understanding* (ASRU), IEEE Workshop. 535-540.

Wise, B., Cole, R., Van V, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., et al., (2005). Learning to read with a virtual tutor: Foundations to literacy. *Interactive literacy education: Facilitating literacy environments through technology*, 31-75.

Wothke, K. (1993). Morphologically based automatic phonetic transcription. *IBM systems Journal*, *32*, 486-511.

Yang, H., Oehlke, C., & Meinel, C. (2011). German speech recognition: A solution for the analysis and processing of lecture recordings. In *International Conference on Computer and Information Science (ICIS)*, *10*, 201-206.

Yoon, S. Y., Chen, L., & Zechner, K. (2010). Predicting word accuracy for the automatic speech recognition of non-native speech. In *Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, *11*, 773-776.

Yu, K., Gales, M., Wang, L., & Woodland, P. C. (2010). Unsupervised training and directed manual transcription for LVCSR. *Speech Communication*, *52*(7), 652-663.

Yuan, J., & Liberman, M. (2011). Automatic detection of “g-dropping” in American English using forced alignment. In *IEEE Workshop on Automatic Speech Recognition & Understanding*, 490-493.

Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., & Wang, W. (2013). Automatic phonetic segmentation using boundary models. In *Interspeech Annual Conference of the International Speech Communication Association*. 2306-2310.

Zekveld, A. A., Kramer, S. E., Kessens, J. M., Vlaming, M. S., & Houtgast, T. (2008). The benefit obtained from visually displayed text from an automatic speech recognizer during listening to speech presented in noise. *Ear and hearing*, *29*(6), 838-852.