AUTOMATIC SPEECH RECOGNITION SYSTEM FOR PEOPLE WITH SPEECH DISORDERS

by

MANTHIBA ELIZABETH RAMABOKA

DISSERTATION

Submitted in fulfilment of the requirements of the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

in the

FACULTY OF SCIENCE AND AGRICULTURE

(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Mr MJD Manamela

COSUPERVISOR: Dr N Gasela

DEDICATION

To my beloved family

DECLARATION

I declare that **AUTOMATIC SPEECH RECOGNITION SYSTEM FOR PEOPLE WITH SPEECH DISORDERS** is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references and that this work has not been submitted before for any other degree at any other institution.

_____ _____

Manthiba Elizabeth Ramaboka Date

ACKNOWLEDGEMENTS

I want to thank the following people for their respective contributions to this dissertation.

- My parents for their love, support, encouragement, and their sincere belief in me.
- My husband, Chris Letsoalo, for his unconditional love, support and encouragement.
- My two children, Neo and Kegaogetswe, for their support and understanding.
- A special thank you to my supervisors, Mr MJD Manamela and Dr N Gasela, for their guidance, support and encouragement.
- My technical advisors, Mr Mabu Manaileng and Dr Thipe Modipa, for their constant support on this study. They helped me throughout the study with their invaluable original ideas.
- My colleagues and friends in the department of Computer Science and department of Mathematics, Science and Technology Education (DMSTE) for their constant support, assistance, and making me feel at home and most of all, the good times we shared.
- University of Limpopo Telkom Centre of Excellence for Speech Technology (ULTCoE4ST) and National Research Foundation (NRF) for their financial support and resources.
- The University of Limpopo Research Office, University of Limpopo Women's Solidarity Association (ULWASA), Council for Scientific and Industrial Research (CSIR) - Human Language Technology (HLT) and the Centre for High Performance Computing (CHPC) for their productive workshops that left me equipped with discipline specific knowledge and technical capabilities.
- Dr Francis Mavhunga for editing this work.

ABBREVIATIONS AND ACRONYMS

ANN – Artificial Neural Networks

ASR – Automatic Speech Recognition

CMN – Cepstral Mean Normalization

CMVN – Cepstral Mean and Variance Normalization

CSIR – Council for Scientific and Industrial Research

CVN – Cepstral Variance Normalization

DTW – Dynamic Time Warping

HLT – Human Language Technology

HMM – Hidden Markov Model

HTK – Hidden Markov Model Toolkit

LPC – Linear Predictive Coefficient

MFCC – Mel Frequency Cepstral Coefficient

MLF – Master Label File

MLP - Multi-Layer Perceptron

PLP – Perceptual Linear Predictive

RBF-Radial Basis Function

TCoE4ST – Telkom Centre ofxcellence for Speech Technology

WER – Word Error Rate

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

ABSTRACT

The conversion of speech to text is essential for communication between speech and visually impaired people. The focus of this study was to develop and evaluate an ASR baseline system designed for normal speech to correct speech disorders. Normal and disordered speech data were sourced from Lwazi project and UCLASS, respectively. The normal speech data was used to train the ASR system. Disordered speech was used to evaluate performance of the system. Features were extracted using the Mel-frequency cepstral coefficients (MFCCs) method in the processing stage. The cepstral mean combined variance normalization (CMVN) was applied to normalise the features. A third-order language model was trained using the SRI Language Modelling (SRILM) toolkit. A recognition accuracy of 65.58% was obtained. The refinement approach is then applied in the recognised utterance to remove the repetitions from stuttered speech. The approach showed that 86% of repeated words in stutter can be removed to yield an improved hypothesized text output. Further refinement of the post-processing module ASR is likely to achieve a near 100% correction of stuttering speech

Keywords: Automatic speech recognition (ASR), speech disorder, stuttering

1. INTRODUCTION

1.1 Overview

Automatic Speech Recognition (ASR) is an audio digital technology that allows a computer to recognise spoken words and converts them to equivalent text. An ASR system receives a speech waveform as an input for processing and produces a sequence of words as an output (Rabiner and Juang, 2004). Compared to other input modes, such as a keyboard, a mouse and other computer access points, the speech input rate is faster when using this technology. Furthermore, in using this technology a user can use a natural speaking style to interact with a speech enabled system.

Today, there are a number of commercial ASR systems on the market, e.g., Google Voice Search, Voice dialling systems, etc. Many people with and without disabilities are benefiting from the development of these speech-to-text conversion systems. They find use in telephone-directory assistance to search for phone numbers, in spoken database querying to search for information, in medical applications to retrieve medical documents. They also used in office dictation to draft and edit documents, and in pronunciation for automatic voice translation into foreign languages (Young, V., and Mihailidis, 2010). As one of the enabling technologies, ASR systems are commonly recommended for medical doctors as assistive tools in the automatic recognition and detection of speech disorders (Van Nuffelen et al. 2008).

A speech disorder is a type of communication disorder which affects the manner of speaking of individuals (Ruben R. J., 2000). It may also affect the production of a voice, the pitch, loudness, type and quality of speech. Many people with speech disorders do not have any problem in reasoning capacity problems or understanding the use of languages. For example, a person with cerebral palsy may have a speech disorder but have no difficulties in understanding spoken languages used for communicating or processing ideas. (http://www.asha.org/public/speech/disorders)

Stuttering, also known as stammering is one of the most common types of speech disorder that affects many people, especially at a young age. This type of disorder is often associated with the repetition of syllables and words or parts of words and struggle to get words out (Ravikumar, et al., 2009; Nöth et al., (2000). The majority of stuttering disorder can be classified as follows:

- Filled pauses *"uh", "um", "eh"*
- Repetitions "the the the",
- Broken words (Words that are not completely pronounced)
- Prolonged sounds (Sounds judged to be improperly prolonged)

In many cases, the exact cause of these speech disorders is unknown (Hollingshead and Heeman, 2004). However, speech disorders can be caused by hearing loss, brain injury, mental illness, alcohol abuse or drug abuse. According to the American Speech Language Hearing Association (ASHA), many types of these speech disorders can be treated by means of speech therapy, but others require medical attention by phoniatric doctors.

1.2 Research Problem

Generally, ASR systems for speech disorder takes a speech signal as an input and converts it to text. Such systems are used by speech pathologists in identifying if a person has a speech disorder. However, the system does not adequately assist people who stutter because it produces text that has repetitions just as in the speech. An algorithm is needed to identify and remove the repetitious text which would otherwise make the output superfluous and meaningless. For a correct recognition, a word is not expected to repeat after another. It is syntactically and grammatically incorrect for identical words to follow each other in the same sentence. If in the same sentence, the same word appears twice or more in succession, then only one word is retained and the rest are discarded. Therefore the algorithm will recognise and delete the repetitions.

This is important so that the combination of ASR and the algorithm will work the same as normal speech recognition for other functions. Output from such a system would be free of spelling and syntactical errors. This ensure accuracy of transcription and utility in business and ordinary communication.

1.3 Purpose of the Study

To develop an ASR baseline system and incorporate an algorithm to recognise repeated words in disordered speech. An algorithm will produce corresponding text without repetitions in the output. Therefore, the objectives of this study were to:

- Source normal and disordered speech data.
- Train an ASR system with normal speech.
- Test a trained ASR system for recognition of normal and disordered speech.
- Evaluate the recognition accuracy of the ASR system.
- Evaluate the output of the ASR system incorporating an algorithm for removing repeated words.

1.4 Significance of the Proposed Study

This study is part of the broader National Collaborative Speech Technology Research Projects that aim at developing speech recognition and synthesis systems for human-machine interaction using the eleven official languages of South Africa (Barnard et al., 2009; Badenhorst et al., 2011). Individuals who are willing to learn reading a new language can do so with the aid of a well-trained ASR system by viewing the text produced by speech recognition in that language.

People with speech disorders and visually challenged people can benefit enormously from ASR system as writing tools. The ASR systems will not only benefit visually challenged people and people with speech disorders, any computing end-user, can use such an ASR system to quickly write their emails just by dictating into the system. This will save people’s time, and reduce the need to constantly focus on computer screens.

1.5 Organization of the dissertation

This dissertation is divided into five chapters. **Chapter 2** introduces the general overview of ASR system, discusses the classifications, as well as the techniques used in ASR system development, and then conclude by reviewing some previous work related to speech disorders. **Chapter 3** presents the research design and methods used in the training and recognition stages. **Chapter 4** displays and discusses the experimental results obtained in this research study. **Chapter 5** presents the conclusions and the recommendations for future research work.

2. LITEARTURE REVIEW

2.1 Introduction

Computer-based processing of speech disorders is a growing innovation found in many fields such as speech therapy, psychology and acoustic signal processing (Kitzing, Maier and Lyberg, 2009). Although ASR systems have been developed for disordered speech, factors such as an individual speaking style, speaking mode and vocabulary size decrease speech recognition accuracy (Husni and Jamaludin, 2010). This chapter gives a historical background and the current state of ASR technology. It also discuss the components of an ASR system, then explore different techniques involved in speech recognition and review some previous work related to the automatic processing of speech disorders.

2.2 Historical background and current state of ASR

Historically, the ASR systems appeared in the middle of the last century. They were not very successful in the beginning. Their performance was limited to the recognition of single words like isolated digits (Kitzing et al., 2009). In 1930 AT & T Bell laboratories developed a speech recognition device with an objective of creating a machine that can imitate human behavior, with respect to speaking naturally and responding properly to spoken language (Juang and Rabiner, 2004). With developing computer technology speech recognition technology also evolved quickly.

ASR systems are currently used in smart phones to send text and email messages. They are also found on the internet to find information, and in automobiles to search for directions. Significant progress has been recorded in the ASR technology in the past decade. However, there are still technological barriers to flexible solutions. ASR systems are still not easily used by people with speech disorders such as stuttering, because of the type and quality of speech input that these people render for processing (Young, V. and Mihailidis, 2010).

2.3 Components of ASR system

The function of ASR system is to receive speech waveform input for processing before it produces a string of words corresponding to it (Rabiner and Juang, 2004). Basically, an ASR is the process by which a computer is able to recognize spoken utterance. Figure 2.1 illustrates the typical block diagram of a speech recognition system. During recognition, the system receives the raw speech as an input. The signal processing component removes noises from the speech input and also reduces the data rate speech signals so that speech can be easily extracted by the system (Huang, Acero and Hon, 2001).

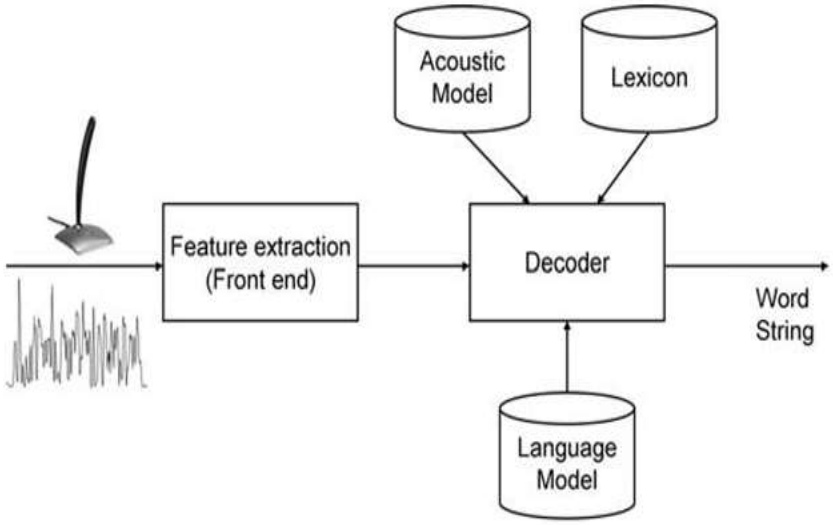

Figure 2-1 A typical block diagram of a speech recognition process (Huang el al. 2001)

As depicted above in Figure 2.1, the *feature extraction* is the first stage of the ASR process. This part receives the raw speech waveform as an input and generates acoustic features. The primary goal of feature extraction is to extract the most relevant information given the speech waveform and to discard as much redundant information as possible (Rabiner, 2004). Various methods are available for efficient extraction of speech parameters such as Linear Prediction Coding (LPC), Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) (Young, S. et al., 2006; Diehl, 2008). The MFCC is the most widely applied acoustic parameterization method for the development of ASR systems.

Acoustic model uses the speech waveforms to generate a sequence of symbols and then compares them to a set of words in the pronunciation dictionary/lexicon. *Lexicon* is a mapping of words to their corresponding pronunciation forms in terms of the phonemes/allophones in a specific natural language. The acoustic model plays an important role during training and decoding phase. For the given acoustic observation X, the goal of speech recognition is to find the most probable word sequence that has the maximum posterior probability P (W|X), where X represents the acoustic features of the word W, P (W) is the language model (Young S, 2008).

The *language model* contains a set of rules for language that is used as the primary context for recognizing words. Language model is used to recognize speech. It consists of a list of words and the probability of their occurrence. This component works closely with the lexicon during recognition to restrict search by limiting the number of possible words that need to be considered in each search. The language model assigns probability to a sequence of words by means of the probability distribution, using the n-grams. It contains a set of rules for a language that is used as the primary context for recognizing words. It also captures regularities in the language. The complexity of the language is directly related to the size of the data on which it is trained. When developing the language model, the training data from a specific domain together with a dictionary are used to create appropriate vocabulary. This ensures that the percentage of the out-of-vocabulary (OOV) word error rate (WER) is decreased. The WER affects the quality of the system and at times, can lead to less recognition accuracy (Huang *et al.* 2001).

The role of *decoding* is to find the solution to the search problem, using a pattern matching breadth-first search algorithm such as the Viterbi decoding algorithm (Huang et al., 2001; Young S, 2008). Decoding process combines the Acoustic Models (AM), Language Models (LM) and the lexicon, and use them to perform the actual recognition process.

2.4 Classifications of ASR systems

Generally, there are three basic types of ASR systems: speaker-dependent, speaker-independent and speaker-adaptive system (Huang *et. al.*, 2001).

The *speaker-dependent* recognition systems are designed to recognise a single user. They are generally more accurate for the particular speaker, but much less accurate with other speakers. Speaker-dependent recognition systems use a template-matching technique. That is, the user's speech is compared to his/her own patterns in the stored reference templates in order to recognize the pattern. The stored reference templates could be either in phonemes or words. Such systems are easy and more accurate to develop, but the more the vocabulary needed for a particular system, the more training required (Young and Alex Mihailidis, 2010).

The *speaker-independent* recognition systems are designed to recognise multiple users. These systems do not require speaker training prior to use. Unlike speaker-dependent recognition system, the speaker-independent system depend on templates created from a multiple number of speakers. Many different speakers are able to use such an ASR system with relatively good recognition accuracy if their speech falls within the range of the collected speech samples. The downside of the speaker independent recognition system is that it is not flexible. The recognition accuracy may be very low for people whom their speech samples are significantly different from the stored templates (Wiśniewski, Kuniszyk-Jóźkowiak, Smołka, and Suszyński, 2007).

The *speaker-adaptive* recognition systems are designed to adjust to a new user without the need to train every word in the system's vocabulary. Like speaker-dependent systems, adaptive-recognition systems are fitted with acoustic templates, so the system requires less training. However, unlike speaker-independent recognition system, as the person uses this system, it constantly updates the acoustic templates with the individual user's speech. Recognition accuracy can be improved as the user continues to use such a system. However, this adaptation may

require supervision to make sure that the system does not “adapt” to wrong word (Young and Mihailidis, 2010).

Besides being speaker-dependent, speaker-independent or speaker-adaptive, ASR systems may also differ by the type of speech they receive and process; discrete speech recognition systems and continuous speech recognition systems (Huang *et. al.*, 1993; Whittaker *et. al.*, 2001).

Discrete speech recognition systems are designed to recognise one word at time. This system require that the users insert explicit pauses between words .The insertion of pauses is to make sure that the word boundaries are adequately recognised. It does not mean that it accepts single words only, but does require a single utterance at a time. Such systems may hold advantages for situation where the user is required to give only one word responses or commands. The disadvantage of this type of recognizers is that, the insertion of pauses may feel and look unnatural and strenuous for some users (Whittaker *et. al.*, 2001).

Continuous speech recognition systems are designed to recognize several words. It does not require the speaker to pause between words. The advantage of continuous speech recognition is that, the input rate is faster and a user can speak naturally. Continuous speech recognizers are most difficult to create because it is not easy for the computer to accurately detect and identify words boundaries between phrases. In addition, it may be more difficult for users with speech disorders such as stuttering to maintain breath support at the sentence level (Young V and Mihailidis, 2010). This decreased intelligibility may negatively affect the speech recognition accuracy.

Another classification can be made according to the size of the vocabulary (Whittaker *et. al.*, 2001). Some ASR applications require numbers only, other systems like the Google voice search require very large wordlists. The size of vocabularies can be classified as follows (Kitzing *et al.*, 2009):

- Small vocabulary – less than two hundreds of words

- Medium vocabulary – two hundreds to five thousand words
- Large vocabulary - greater than five thousand words

In general, the smaller size of the vocabulary, the better the recognition accuracy and the less training required. However, the tasks that can be performed with small vocabularies are more restricted than the large vocabulary systems (Huang *et. al.*, 1993; Whittaker *et. al.*, 2001). Furthermore, the classification can be made according to the size of their linguistic recognition units whether word-based or phoneme-based speech recognition systems (Huang *et. al.*, 2001).

A *word-based* speech recognition is a system in which the smallest recognition unit is a word. The recognition accuracy is very high because the system is free from negative side effects of co-articulation. However, for continuous speech recognition, transition effects between words may cause problems. Furthermore, the processing time and memory requirements are very high because there are many words in a language which are the basis of the reference patterns.

Phoneme-based speech recognition is a system in which the recognition units are phonemes. While recognition accuracy decreases in a phoneme-based system, it is possible to apply error detection and error-correction using the ability to produce fast results with very few phoneme numbers. There can be several speech recognition systems that make use of sub-word units like diphone-based, triphone-based, and syllable-based.

There are other factors that affect speech recognition such as environment variability, low microphone quality, microphone-to-mouth distance, sex, age, social background, personal physical or emotional state and speed of speech. Despite these difficulties, the development of ASR systems has been a great success in number of applications (Wiśniewski et al., 2007). There has been a growing interest in speech–to-text technology in speech therapy for recognition and of speech disorders and research (Young and Mihaildis, 2009; Czyzewski et al., 2003). Today ASR systems have become more speaker-independent and accept large vocabularies and continuous speech (Wiśniewski *et al.*, 2007).

Several factors have played a major role in the development of ASR systems. One key factor is the introduction of a more effective statistical algorithm, the hidden Markov models (HMMs) technique (Wiśniewski *et al.*, 2007). Other techniques are template matching, artificial neural networks (ANN), knowledge based system and acoustic phonetic. These above technologies vary in speed, accuracy and storage requirements. The *HMM* is a stochastic model that captures the statistical properties of observed real world data (Tian-Swee, Helbin, Ariff, Chee-Ming, and Salleh, 2007). This technique is most widely used in the recognition of disorders such as speech dysfluency: repetition and prolongation. It uses two transitions between states to quickly search through a database. Each state represents part of a speech unit and contains a probability distribution that describes the acoustic properties of a speech frame. The transitions are associated with probabilities of proceeding from one state to another when the computer is trying to match spoken input with the stored model. Hence no direct matching between stored models and input is involved. Transitions may proceed from one state to the next, may skip a state, or they may call recursive and return to the previous state. These processing steps depend much on adequate language models, both at the phoneme, word and sentence level (Kitzing *et al.*, 2009).

The ANNs are mathematical model or computational models that try to develop intelligent systems (Ravikumar *et al.*, 2009, Husni and Jamuldin, 2010). This technique is inspired by the functionality of the human brain. Researchers from many scientific disciplines are designing ANNs to solve problems in pattern recognition, prediction, optimization, associative memory and control. ANN technique functions at a phoneme recognition level. A speech frame is received by the input layer. The network detects information correlated to phonemic features and refines and stores this information. Each node in a higher layer sums the signal that it receives from the connected lower level nodes in the network after amplifying each input by a weight that reflects the importance of that connection. After passing the sum through a threshold function, the network constitutes the output of that node. The output layer then communicates the hypothesis as to what the identity of the current speech frame is to the external environment. The ANNs play an

important role in both speech and they are capable of solving much more complicated recognition task. In recent years, ANNs are widely used in many ways in speech disorders, such as recognition of dysfluency in stuttered speech, but do not perform better as HMM when it comes to large vocabularies (Ravikumar et al, 2008).

The SVM is a powerful machine learning tool that is widely used in the fields of pattern recognition (Campbel et al., 2006). As noted by Ghai and Singh (2002), SVM optimization problem attempts to obtain a good separation hyper-plane between two classes in the higher dimensional space. To date, SVM was used as classification tool in stuttering recognition (Ravikumar et al, 2009).

Template matching is performed at the word level. The user's speech is compared with stored templates as in Figure 2.2, in order to recognize the pattern. This has the advantage of using perfect word models. However, training is required in template matching. The purpose of training is to provide the computer with enough versions of all spoken utterances to be recognized (Ravikumar et al, 2009). The selected template is called the best match for the input.

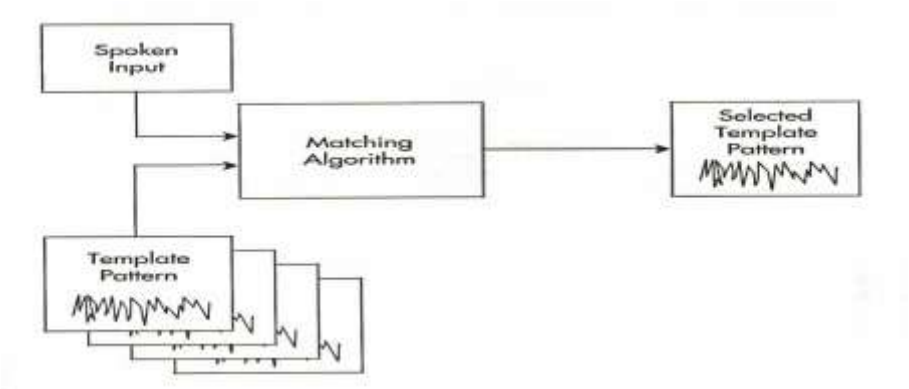

Figure 2-2 Recognition Using Template Matching (Ravikumar et al., 2009)

Template matching performs very well with small vocabularies of phonetically distinct items but has difficulty making fine distinctions required for large vocabulary recognition and recognition of vocabularies containing similar sounding words (called confusable words). Since it operates at the word level there must be at least one stored template for each word in the application vocabulary. If, for example,

there are five thousand words in an application, there would need to be at least five thousand templates (Ravikumar *et al*., 2009).

Knowledge based system uses a set of features from the speech waveform, and the training system generates set of production rules automatically from the speech samples. These rules are derived from the parameters that provide most information about a classification. The recognition is performed at the frame level, using an interface engine to execute the decision tree and classify the firing of the rules. This has the advantage of explicitly modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully (O'Brien, 1993).

2.5 Related Studies on Speech Disorders

Many studies for automatic recognition of speech disorders have been done (Chee, Chia and Sazali, 2009). This section presents an overview of previous works found in the literature. It concentrates on how ASR is being performed, designed. It also focuses on how the experiments are analysed.

Ravikumar et al. (2008) proposed an automatic detection method for syllable repetition. The detection scheme was divided into four stages; segmentation, feature extraction, score matching and decision logic. The feature were extracted using MFCC. The recognition system was based on ANN. The speech data consists of ten samples. 80% of speech samples were used to train the system and the remaining 20% were used for testing. The recognition accuracy of 83% was achieved. They (Ravikumar *et al.*, 2009), further proposed another detection method for stuttered dysfluencies using MFCC and SVM (Reda and Khoribi, 2008). Fifteen speech samples were collected. Twelve samples were used for training and the remaining three samples were used for testing. The best word recognition accuracy of 94.35% was achieved. The SVM performed much better with average results of 94.35% when compared ANN.

Wiśniewski et al. (2007) presented two papers about an automatic detection of speech disorders in continuous speech using the HMM technique. The HMM is a

stochastic model that is widely used in speech recognition system (Al-Alaoui et al., 2008). In the first paper, they employed thirty-eight samples for prolongative of fricatives recognition model, thirty samples for stops blockade of recognition model and thirty samples for summary model. All the recordings were normalized to the same dynamic range. The language for implementation was Polish-an official language that is used throughout Poland. The best phone recognition accuracy of 70% as. In the second paper they (Wiśniewski *et al.,* 2007) proposed an automatic detection of prolonged fricative phones with HMM as classification technique. The frequency of the sound samples was 22050 Hz. All records were normalized to the same dynamic range -50dB.The best phone recognition results of approximately 80% were achieved.

Tian-Swee *et al.* (2007) evaluated an automatic stuttering recognition system using HMM technique. The database consisted of twenty samples of normal speech data and fifteen samples of artificial stutter speech data. Ten samples of each normal and artificial stutter were used to generate a speech model set. Remaining five samples of normal speech data and artificial stuttered speech data were used to test on HMM models. The normal speakers achieved an average recognition accuracy of 96%. The artificial stutter speech achieved a recognition accuracy of 90%. Despite yielding a high accuracy as 90%, the accuracy was being questioned due to the reason that system was not tested on speech samples taken from actual stuttering clients.

Świetlicka et al. (2009) presented an automatic detection of dysfluency in stuttered speech. They employed eight stuttering people for speech recordings. Fifty- nine fluent speech samples and fifty- nine non-fluent speech samples were obtained from the recordings. The parameters of the speech samples were used as an input for the networks. They applied Multilayer Perceptron (MLP) and Radial Basis Function (RBF) networks to recognize and classify fluent and non-fluent speech samples. MLP is a feature extraction technique that can be given multiple frames of input features, by enabling modeling of a larger temporal window (Stolcke et al., 2006). An average percentage of word recognition accuracy of 91.5% was achieved for all networks.

Although the best recognition results were obtained in all these studies related to recognition of speech disorders, there were some comments regarding the consistency of the results due to amount of training speech data used. For example, Ravikumar et al. (2009 and 2008), used only eight speech samples to train the system and it was tested with only two speech samples. Although they achieved 83% of recognition accuracy, the accuracy was being questioned and it was suggested that the amount of training data should be increased so that the system can capture all acoustic and prosodic models of stuttered speech. In addition, the speech databases in findings from (Ravikumar et al., 2009) were created without taking into consideration the distribution of features such as gender, age or origin. Moreover, some of the systems were tested with artificial or simulated stuttered speech.

In 2006 and 2010, Maier and Schuster investigated the applicability of the method to speech disorders caused by head and neck cancer. Intelligibility was quantified by speech recognition on recordings of a standard text read by forty-one German *laryngectomized* patients with cancer of the larynx and forty-nine German patients who had suffered from oral cancer. All the patients read a phonetically rich text with 108 words. The data were recorded using a close-talk microphone with 16 kHz sampling frequency and 16 bit resolution. The ASR system based on HMM was used. About 80% of the 578 training speakers were between 20 and 29 year old, less than 10% were over forty years old. The ASR system was integrated into the "Program for the Evaluation and Analysis of all Kinds of Speech disorders" (PEAKS). The non-adapted ASR system for automatic speech evaluation that has previously been proven to be adequate for "normal" speech samples was applied. The automatic speech evaluations were compared to a control group of forty speakers without speech pathology. An increased age has been shown to have a negative influence on automatic speech recognition. The word recognition rate of 76% on average was obtained. The results of the control group demonstrated that the standard deviation in word recognition rate of normal speech in speakers of the same age is about half of the pathologic one. It was concluded that the normal data

for all age class could quantify a patient's intelligibility in relation to norm in percent ranks.

Husni et al. (2010) reported about ASR performance using context-dependent phoneme models. The vocabulary consisted of 114 words was used for recordings. The words contained all syllable patterns (consonant-vocal pair) that make up valid words in Bahasa Melayu. Bahasa Melayu is the first language in Malaysia. Participants were ten dyslexic children, from 7-14 years old whose reading levels were similar.114 words were prompted randomly and the participants were required to read aloud each of the word into a head-mounted microphone. The hybrid artificial neural network (ANN) was the chosen as the training method for the performance. The recognition accuracy 75% was obtained when using context-dependent (CD) phoneme model and phoneme refinement rule.

Shunsuke et al. (2011) developed a speech support system using body-conducted speech recognition. To develop the system, they first constructed a signal database for healthy individuals and patients with speech disorders. Speech data was collected from three Japanese, One normal speaker and two speakers with speech disorder respectively. Table 2.1 shows the conditions of speakers. One patient had a damaged pharynx and was examined before surgery. The other patient was examined after surgery.

Table 2-1: Statistics of healthy individuals and patients with speech disorders

User type	Condition
Healthy	22 years old male;
	No damaged vocal chord.
Disorder 1	28 years old male;
	Own damaged vocal chord with polyp.
Disorder 1	61 years old male;
	Speech reconstruction.

Shunsuke et al. (2011)

The difference between the speech of the healthy person and the patients using recognition parameters with the database was analysed. The system used feature vector and HMM technique to calculate likelihoods. The system performance was evaluated by considering word correct rates and word accuracy rates. Both

experiments revealed a word recognition rate of around 60% for the speech of the healthy subject. However, a word recognition rate of only 30% to 40% was found for the body-conducted speech in patients with speech disorder, because of its low quality signal. The recognition results indicated that it is difficult for the speech recognition system to function accurately in the presence of speech disorders. Even if body-conducted speech recognition for speech disorders is used in speech recognition, good recognition performance cannot be obtained. To achieve sufficient speech recognition performance, the system requires the estimation of new acoustic models using speech disorders, although there is no experimental evidence for this hypothesis.

Muhammad, et al. (2011) investigated the accuracy of the conventional ASR system. Features were extracted using MFCC technique. Training was implemented using HMM technique. Two recognition systems were developed. The first system was trained using forty normal speech and tested with 10 samples of normal speech. The recognition accuracy of 100% was achieved. The other speech recognition system was trained using normal speech and tested with speech disorders from sixty-two dysphonic patients. The recognition accuracy varied between 56% and 82.50%. All participants were Arabs. Their age ranged from 18-50 years. The recorded speech data were isolated Arabic digits from one to ten.The results revealed that, there was a significant loss of recognition accuracy when the system was evaluated using speech disorder. It was concluded that the current ASR technique is far from reliability in recognizing the speech disorders.

Salama, Khoribi and Shoman (2014) presented a paper about ASR for people with dysarthria speech disorders based on both speech and visual components. In their work, the focus was on adding Discrete Cosine Transform Coefficients of mouth region as visual features for people with speech disorders. The database consists of five tasks of connected words produced by individual with speech disorders was used to evaluate the performance of the ASR system. The tasks contained ten digits (zero-nine), twenty-six alphabet words, nineteen computer command, hundred common words and hundred uncommon words. The hidden Markov models (HMM) was the chosen training and testing method for the performance. It was found that

the visual features are highly effective and the recognition accuracy was increased by 7.91% in speaker dependent system and 3% for speaker independent system.

Christensen et al. (2014) investigated the recognition accuracy of disordered speech by selecting which speakers to include in the speaker independent model. The speech database from the Universal Access (UA-Speech) was used for training and testing of acoustic models. The database consists of speech from fifteen speakers with a range of impairment levels. All acoustic models were based on the HTK toolkit. No language model was used, and the decoding was done having all possible testing words in parallel in addition to silence models at the start and end. There was an improvement of 11.5% on the accuracy.

Yunbin et al. (2009) presented a pioneering effort on recognition of speech disorders using acoustic features and surface electromyographic (sEMG) signals the feature extraction was implemented using MFCC with CMVN to normalise sEMG signals. The recognition was based on HMM technique. The speech data consists of eight healthy English speakers was used to train the system. The other speech data from five speakers with speech disorders was used to test the system. One speaker had suffered a stroke. Four speakers had speech impairment caused by cerebral palsy. Due to limited vocabulary of isolated words, the speech data consisted of eleven digits, twenty-six alphabets and words like yes, no, left and right was used. A headset microphone one was used to collect acoustic signals and it was positioned approximately five centimetre in front of the mouth while sEMG were collected using eleven sEMG sensors. The average recognition accuracy of 54% was achieved. Results indicated that speaker dependent isolated-word recognition was highly. It was concluded that further development of speech recognition systems using sEMG is needed to increase usability and robustness.

.

In their work, the focus was on adding Discrete Cosine Transform Coefficients of mouth region as visual features for people with speech disorders. The database consists of five tasks of connected words produced by individual with speech disorders was used to evaluate the performance of the ASR system. The tasks contained ten digits (zero-nine), twenty-six alphabet words, nineteen computer

command, hundred common words and hundred uncommon words. The hidden Markov models (HMM) was the chosen training and testing method for the performance. It was found that the visual features are highly effective and the recognition accuracy was increased by 7.91% in speaker dependent system and 3% for speaker independent system.

2.6 Conclusion

In this chapter, the historical background and the current state of ASR technology were discussed, a general overview of an ASR system including classification of, and techniques used in speech recognition systems were also discussed. A review on previous work related to speech disorders was also presented, where the use of HMMs and Mel-frequency cepstral coefficients (MFCC) techniques have proven to give highest recognition accuracy of 96%. Although previous work has shown a potential effort on speech recognition accuracy of disordered speech, the effectiveness of using MFCC and HMM has shown in using small amount of data.

3. RESEARCH METHODOLOGY

3.1 Introduction

The objective of this research study was to develop an ASR baseline system and apply an approach to enhance the quality and surface representation of the recognized utterances. This chapter gives an overview of the proposed approach. The approach involves training a standard HMM-based recogniser using normal speech and testing it with disordered speech. The details are discussed in the succeeding sections.

3.2 Overview of an ASR System Development Approach

An ASR system incorporating an algorithm was trained with normal speech records and tested with disordered speech. Features were extracted using the Mel-frequency cepstral coefficients (MFCCs) method in the processing stage and thereafter applied the cepstral mean combined with variance normalization (CMVN). A third-order language model (LM) was trained using the Stanford Research Institute Language Modelling (SRILM) toolkit. The phonemes were modelled by a 3-state left-right HMM. Sixteen Gaussian mixture models were employed as state-conditioned output probability distributions HDecode tool was used to evaluate the recognition performance of the speech recognition system. All the software packages used in this research project are available for free downloading at their respective websites.

General system development approach is given Figure 3.1.

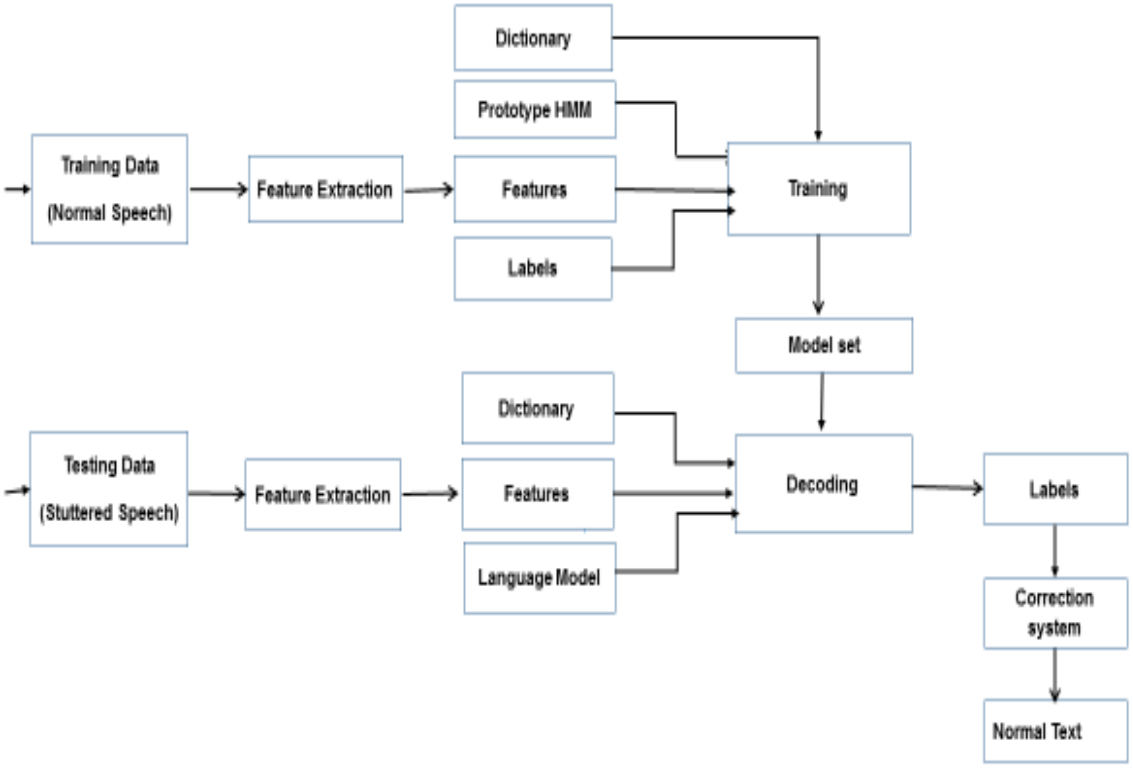

Figure 3-1: An overview of the ASR system

As outlined in Figure 3.1, the pronunciation dictionary, language model and acoustic features of the disordered speech data were used for decoding the models. The recognised utterances were further passed to a correction algorithm for post-processing. The algorithm automatically corrects, syntactically, the recognised hypothesis. The next sections explain the research process in details.

3.2.1 Data Preparation

The development of any speech recognition engine involves collection of the sufficient and appropriate speech data. Language-specific speech data is essential for system training and testing purposes. In this study, the English training speech data used was obtained from the Lwazi speech project (Meraka-Institute. Lwazi ASR corpus. 2009).The speech data was developed by the Human Language

Technology (HLT) research group at the Council for Scientific and Industrial Research (CSIR) (Van Heerden et al., 2009). The data is freely available from the Resource Management Agency (RMA)**.** It consists of 5843 phonetically rich utterances made by 200 speakers (Modipa & Davel, 2012). Orthographic transcriptions are also provided in the same speech data.

For testing, the English speech data was obtained from the University College London Archive of Stuttered Speech (UCLASS) database. The speech data was chosen as it is one of the largest database available with 15 speakers. The speech corpus is designed for multiple research purposes which include behavioral investigations, comparison of dysfluent behaviors, clinical applications and machine recognition. For these purposes, manual orthographic transcriptions at word level are provided. From the database, 149 utterances from 5 different speakers were selected. The speakers were selected because their utterances were balanced in terms of dysfluencies that occur in stuttering. The speakers were aged 11-20 years who were referred to clinics in London for assessment of stuttering. The speech data was used to establish the acoustic properties of stuttered speech. Table 3.1 shows the speaker age and conditions in the testing data.

Table 3-1: Speaker age and conditions

	Age(years)	**Types of disfluencies**
Speaker 1	20	Repetitions and filled pauses
Speaker 2	13	Broken words and prolonged sounds
Speaker 3	11	Repetitions, filled pauses, broken words and prolonged sounds
Speaker 4	13	Repetitions and filled pauses
Speaker 5	20	Repetitions and filled pauses

The amount of speech data that was used for the ASR system development is summarized in the table below.

***Table* 3-2: Amount of training and testing speech data**

	Training set	Testing set
Number of speakers	196	5
No. of utterances	5843	149
Number words	40908	1340
Duration (hours)	17	0.21

3.2.2 The Language model

There are several software packages for statistical language modelling that have been in use for the past decades. In this study, a freely available Stanford Research Institute Language Modeling (SRILM) toolkit and the Lwazi orthographic transcriptions were used to build trigram Language Models (LMs) for training. SRILM is a toolkit for producing statistical LMs, primarily for use in speech recognition. The toolkit supports the development and estimation of a range of LM types based on n-gram statistics. To accomplish these two purposes, the SRILM uses the tools, *ngram-count* and *ngram,* respectively. The main parameter controlling the evaluation of LMs is the *-ppl* which measures the perplexity (Stolcke, 2002). The following command was executed to train a trigram LMs:

> *ngram-count*-text train.txt-*order 3 -lm trigram.lm –interpolate -cdiscount1 0.5 cdiscount2 0.5 -cdiscount3 0.5.*

The *ngram-count* command estimates the word probabilities from the training data. The *train.txt* file is the training data which contains all words from the sentences in transcriptions of the training data set. An interpolated absolute discounting with a discounting coefficient of 0.5 was used. As shown by Whittaker & Woodland, (2001),
the LM order, like the discounting coefficient, can be specified arbitrarily by the researcher.

A sample of the trained LMs is outlined in Fig 3.2.

```
\data\
ngram 1=1995
ngram 2=5834
ngram 3=4252

\1-grams:
-0.9027703</s>
-99     <s>     -1.41007
-1.838073       a       -1.056197
-3.62577        abandoned       -1.255273
…
\2-grams:
-1.927614       <s> a   -0.6670728
-3.130765       <s> absurdities -0.9030899
-2.203001       <s> after       -0.9435346
…
\3-grams:
-1.318106       <s> a c
-0.6614732      <s> a cell
-0.7974948      <s> a heavy
…
\end\
```

Error! No sequence specified.**Figure 3-2: A sample of the trained statistical language model**

To build LM testing data, all the testing transcriptions were extracted. A file test.txt was generated from the sentences in the testing data set. The trained LMs is then evaluated on the test data by executing the *ngram* command with a *–ppl* parameter as follows:

ngram -ppl test.txt -order 3 -lm trigram.lm

Table 3.3, shows the summary of the trained LMs including the total words in the LM, the total trigrams, number of out of vocabulary (OOV) words and the perplexity of the test set

Table 3-3: Summary of the language models

Total sentences	4756
Total words	33264
Total trigrams	4292
OOVs	123

Test Set Perplexity	6.22

3.2.3 The Pronunciation Dictionary

The pronunciation dictionary can be thought of as a mapping of words to their corresponding pronunciations. It is different from a standard dictionary in that, it does not provide the meanings of the words but only how they are pronounced. In a typical ASR, a pronunciation dictionary contains generic words whose pronunciations are represented by phonemes.

Building an accurately hand-crafted pronunciation dictionary for a language can be a very difficult task. It is typically created manually by linguistic experts of the target language. Depending on the size of the vocabulary, the manual approach can be time consuming, expensive and therefore an undesirable option. Several methods of creating a dictionary have been introduced in the past two decades (Black et al., 1998 and Besling, 1994). Most of these approaches are based on finding rules for the conversion of the written word to a phonetic transcription. This can be either by applying rules (Black et al., 1998) or by statistical methods (Besling, 1994).

For this research study, the Lwazi English pronunciation dictionary was used. The dictionary was developed to be used for speech technology systems (Davel. & Martirosian, 2009). It consists of approximately 3112 frequently occurring words in South African English. The dictionary was used to train the acoustic models.

However, for the testing process, the dictionary was automatically created using Lwazi grapheme-to-phoneme rules (Davel & Martirosian, 2009). We created a new pronunciation dictionary called *dict*, using transcriptions of the test data. We merged the two dictionaries (Lwazi and *dict)* into one pronunciation dictionary (*baseline.dict*) to accommodate all the missing words in either dictionary. The total number of words in the pronunciation dictionary is listed in Table 3.4.

Table 3-4: Total number of words in the dictionary

Language	Phones	Unique words
English	44	3716

3.2.4 Creating the Transcription File

To train a set of HMMs, every audio file (.wav) must have a corresponding word level transcription file (.txt). The Master Label File (MLF) that contain label of each line was created. The MLF combines the transcriptions into a single file. The script *prompts2mlf* is used on the *words.txt* to generate MLF file, *words.mlf*. A part of MLF is shown below in Figure 3.3.

```
#!MLF!#
"*/english_160_01.lab"
english
.
"*/english_160_02.lab"
twenty
four
.
"*/english_160_03.lab"
female
.
"*/english_160_04.lab"
cell
phone
.
```

Figure 3-3 Error! No sequence specified.**: A sample of the Master Label File (MLF).**

The MLF file can further be expanded into a phone level by applying the *HLEd* command:

HLEd –d dict –i monophones.mlf monophones.led words.mlf

is executed to replace each word with its phonemes and the results are stored in a file named mono*phones.mlf*. The phone level transcriptions produced by the command above, is shown in Figure 3.4. Each phrase is expanded into its phonemes.

```
#! MLF! #
"*/english_160_01.lab"
sil
E
N
g
l
i
S
sil
.
"*/english_160_02.lab"
sil
t
w
E
n
t
i
f
O_c
r_b
sil
.
.
                                        .
```

Figure 3-4: A portion of the phone level transcriptions from the MLF

The *monophones.led* file contains the following lines:

EX
IS sil
DE sp

The EX (expand) replaces each words in *word.mlf* by its corresponding pronunciation. The IS insert a silence model *sil* at the beginning and end of each utterances. Lastly, the DE removes all short pauses (*sp*) which are not needed .

3.2.5 Feature Extraction

The final stage of data preparation is to extract features from speech waveform. Feature extraction plays an important part in both training and recognition phases. Although various methods are available for efficient extraction of speech

parameters, in this study, the MFCCs is chosen as the feature extraction method. The following line specifications and conversion parameters was set up in the configuration file to extract features from raw speech waveforms.

```
#Coding parameters
CEPLIFTER       =       22
ENORMALISE      =       FALSE
NUMCEPS         =       12
NUMCHANS        =       26
PREEMCOEF       =       0.97
SAVECOMPRESSED=         FALSE
SAVEWITHCRC     =       FALSE
SOURCEFORMAT    =       WAVE
TARGETKIND      =       MFCC_0_D_A_Z
TARGETRATE      =       100000.0
USEHAMMING      =       TRUE
WINDOWSIZE      =       250000.0
ZMEANSOURCE     =       TRUE
LOFREQ          =       150
HIFREQ          =       4000
```

Figure 3-5: The configuration file of the MFCC

In brief, features (for both data sets) were extracted with the TARGETKINDS = MFCC_0_D_A_Z as the energy component. The Fast Fourier Transform (FFT) use the Hamming window size of 25ms and is applied with an overlapping frame rate of 10ms. The pre-emphasis coefficient of 0.97 was used to normalize the energy.

The Cepstral Mean and Variance Normalization (CMVN) was used to perform normalization. The CMVN technique is a combination of Cepstral Mean Normalisation (CMN) and Cepstral Variance Normalisation (CVN) (Manaileng and Manamela, 2014). During this stage, a list which has a full path were the MFCCs are stored is created. The configuration file and HTK tool HCopy converts audio files to MFCCs. The MFCCs representation captures all the necessary speech acoustic features extracted from the original speech waveforms. The extracted parameters are used in training and testing.

3.2.6 Training

The first procedure in training the HMM is to define the prototype model. The important part in this procedure is to define the model topology. A Markov prototype

model file called "*proto*" was created, thereafter, the HCompV command was executed to generate a new proto model. The proto model runs with the script file *"train.sh*" under the command *HCompV* to produce the two files, "*vFloors*" and "*proto*". The files were then stored in the folder "*hmm0*". Another file called Macros was created by combining the contents of "*vFloors"* with "*proto*". An HMM definition file "*hmmdef*" was created by adding each phone in *monophones1* with the corresponding prototype. Each phone is put in double quotes preceded by "~h".

The models in *hmm* directories were further re-estimated by running the command line:

HERest –C config –I phones0.mlf –t 240 p–100 –S 10.train.scp –H hmm0/macros –H hmm0/hmmdefs.mmf –M hmm1 monophones0

The values in the command line represent the pruning threshold (-t), the grammar scale factor (-s) and the insertion penalty. The value of the pruning factor is used to reduce the effective size of the search space during recognition. The word insertion penalty is a value added to each token when it transits from the end of one word to the start of the next word. The goal of grammar scale factor in speech recognition is to detect words in the hypothesised sentence that are likely to have been recognized (Young, 2006).

A three state left-to-right HMM was created for each phone for the silence model "*sil*". Extra transitions from states 2 to 4 and from states 4 to 2 were introduced to make the model robust as individual states absorb the various impulsive noises in the training data (Black et al., 1998). Furthermore a one state "*sp*" model to monophone list, which has its emitting state tied to the centre state of the silence model was created. To achieve this a file called "*sil.hed*" with the following line specifications (Figure 3.6) was created. The HHED command (below) was executed twice to tie the model to the *sil* centre state.

HHEd –T 1 –H hmm4/macros –H hmm4/hmmdefs.mmf –M hmm5 sil.hed monophones_sp.lst

```
AT 2 4 0.2 {sil.transP}
AT 4 2 0.2 {sil.transP}
AT 1 3 0.3 {sp.transP}
TI silist {sil.state [3], sp.state [2]}
```

Figure 3-6: Specification of the sil.hed file

Monophones transcriptions were converted to triphones transcriptions. Triphones are a group of three phones formed from a single phone. Each phone is replaced with right neighbour phone, the main phone, and the left neighbour phone. Lastly states of triphones were tied to ensure that all state distribution can be estimated better.

This was done by the following commands:

- Create a file "*mktri.led*" containing the following line specifications:

```
WB sp
WB sil
TC
```

Figure 3-7: Specification of the mktri.led file

We then execute the HLEd command to convert the monophones to triphones:

- *HLEd –n triphones 1 –l (*) –I wintry.mlf mktri.led aligned.*mlf

A sample of triphones is listed below in Figure 3.8.

```
#!MLF!#
"*/english_001_01.lab"
sil
sil-E+N
E-N+g
N-g+l
g-l+i
l-i+S
i-S+sil
sp
sil.
```

Figure 3-8: A portion of triphones file

The most important step in making tied-state triphones is to run the command HDman against the dictionary to generate a new version of the dictionary which consists of words with their pronunciations represented using triphones.

The HHEd command is executed to perform decision tree state tying and the output is saved in a log file for the purpose of threshold tuning. Finally, the re-estimation of the HMMs and creation of a tied list file is performed by running the HERest command twice. The tied list together with the HMMs can now be used to recognize speech.

3.2.7 Evaluation

HDecode tool was used to evaluate the recognition performance using the test data. HDecode is a package that is an add-on to HTK. It is designed for large vocabulary speech recognition systems (LVCSR) tasks and for the system that require trigram LMs. The tool works more or less the same as HVite tool but, it can handle n-gram LMs up to trigrams (Young S., et al., 2006).

To prepare the test data, the SRILM toolkit and the UCLASS transcriptions were used to build trigram LMs. The unseen words from the UCLASS transcriptions were manually added to the baseline dictionary. The new words were modelled using the

same rules and phone sets as the Lwazi English dictionary. The following command was executed to perform decoding:

HDecode -C $DIR_EXP/config/hvite.cfg -A -D -T 1 -V -H hmm_41/hmmDefs.mmf –H hmm_41/macros –J hmm_36 –S audio.lst -t 240 -s 10.0 -p -10 -w trigram.lm -i results.mlf baseline.dict tiedlist.lst

The HDecode tool uses a list of physical models (tiedlist), the LMs (trigram.lm) and the pronunciation dictionary to recognise a set of audio files (the test set). The values of the insertion penalty, grammar scale factor and beam-width pruning threshold were optimally set for decoding. Each test file was recognised and its output was stored in the file called results.mlf. A portion of the results.mlf file is displayed in Figure 3.9.

```
#!MLF!#
"~/mlfs /M_1017_11y8m_1-018.rec"
0 200000 <s> -178.220398
200000 4900000 for -3446.626465
4900000 9100000 there -2953.368408
9100000 20400000 this -8475.755859
20400000 30000000 lying -7315.979492
30000000 35700000 lying -4762.199219
35700000 46800000 lying -8652.533203
46800000 49600000 and -2282.484375
49600000 54500000 and -3874.734375
54500000 59000000 and -3805.945312
59000000 65600000 romans -4903.210938
65600000 66100000 </s> -397.527344.
```

Figure 3-9: A portion of the MLF file

3.2.8 Improving search results

The last step of the recognition phase was to make improvements on the search results. The goal is to make sure that more useful results are obtained. Otherwise, the results will contain errors and redundancies. To remove the speech disorders, the proposed algorithm below was applied.

Open file:

Select every line containing the REC keyword:

For each line:

read every next word after the previous word:

if next word is identical to previous:

then: discard next word

else: proceed

remove all trailing white spaces

print the output to a file

done.

The algorithm reads each line in the recognised file (labels). If in the same sentence the same word appear more than once, it retains the first word and discard the next word. The output is a normal text without any repetitions of words.

Speech disorders like "*uh","um"* and *"eh"* were treated as noise markers They were not modelled acoustically, hence, they did not form part in testing. As a result, a typical sentence uttered as "the the the dog stole uh um stole my my homework", will be recognised as follows: "the dog stole my homework".

Prolonged words and sounds are problematic during decoding since they can be erroneously recognised as other words. This is because there are usually small pauses between the short utterances as the speaker frames a complete word. For example, the prolonged word for "*mom*" can be *"mmooom"*, the short pauses in-between the actual word could cause the recogniser to recognise individual phonemes such as "*m*", "*mo*", "*oo*", "*om*", etc. Fortunately, in this study prolonged sounds and broken words were not modelled in the pronunciation dictionary. Since the experiment was based on word recognition, prolonged sounds and broken words were automatically recognised as words rather individual's phoneme. Therefore they were automatically handle by the HTK-embedded Viterbi search algorithm.

3.3 Conclusion

In this chapter, a detailed information on the overview of the proposed approach is outlined, the procedure followed in the construction of an ASR system using HTK toolkit is discussed, and the process involved in enhancing the output of system is also discussed. The results obtained at the end of the development will be discussed in the next chapter.

4. RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter presents the experimental results obtained for the study. Two ASR baseline systems were developed. The first ASR system developed was trained and tested with normal speech. The second system was trained with normal speech and tested with disordered speech. The performance of the system was evaluated by correctness of word recognition rates and word accuracy rates. The measurements are defined as follows:

$$WordAccuracy = \frac{N-S+D+I}{N} * 100 \quad (1)$$

$$WordCorrectness = \frac{N-D-S}{N} * 100 \quad (2)$$

Where:

- N: is the number of words in a sentence,
- S: an incorrect word was substituted for the correct word,
- D: a correct word was omitted in the recognized sentence,
- I: an extra word was added in the recognized sentence.

4.2 ASR baseline systems

The training phase was performed using 4704 utterances, while testing phase was performed using 1176 utterances. A word recognition accuracy of 64.82% was obtained for normal speech. Figure 4.1 shows the results of word recognition accuracy.

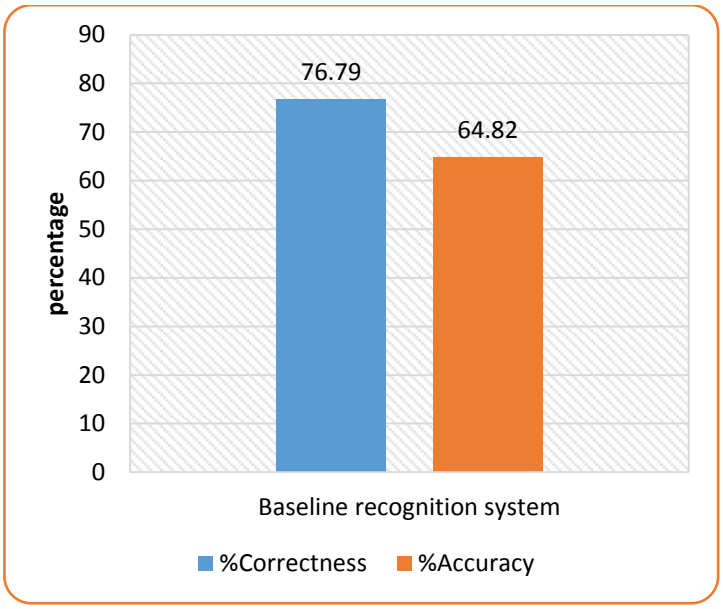

Figure 4-1: ASR system tested with normal speech

The word correctness gives the word level statistics. This indicates that out of the 26743 words in total, 20536 (76.79%) were recognized correctly. The deletion errors (D) were 2429, substitution errors (S) were 3778, and insertion errors (I) were 3202. The recognition accuracy is less than the word correctness, because the accuracy considers the insertion errors.

Figure 4.2 shows the results obtained when the baseline system was tested with disordered speech.

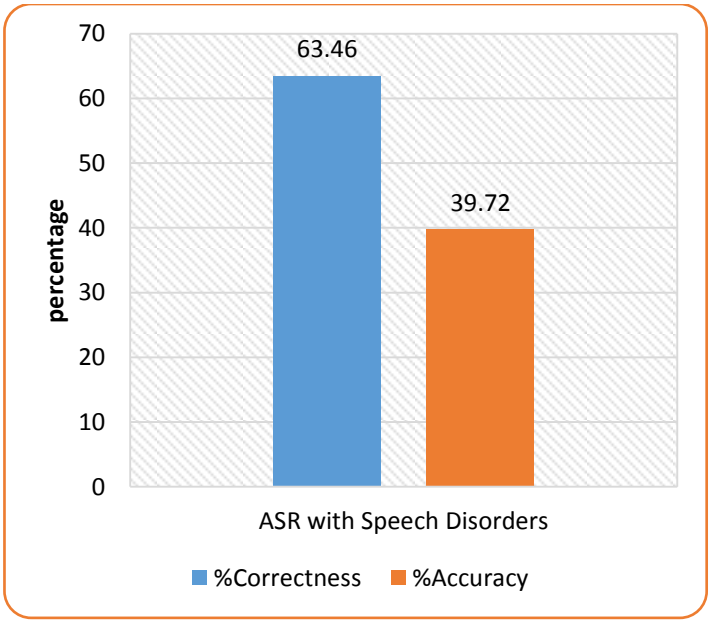

Figure 4-2: ASR tested with disordered speech

From the results reflected in Figure 4.2, the trained ASR system did not perform well with English speech disorders. Both experiments (Figure 4.1 and Figure 4.2) revealed a word correctness of more than 60% for the normal and stuttering speech. However, a decrease in the percentage of recognition accuracy was found when using disordered speech. The recognition accuracy may be affected by many factors such as the instrument used during data collection, size of vocabulary, type and quality of speech and also the surrounding environments (Young V., et al, 2010).

To improve the recognition performance the values of the grammar scale, insertion penalty and the pruning factor were set to a fixed value. These parameters can have a significant effect on the recognition accuracy (Young S. et al., 2006). The next section discusses the steps and process involved in enhancing the recognition performance of the system.

4.2.1 Enhanced ASR System

In order to enhance the recognition accuracy of the recognition system the values of the HDecode parameters were changed to control the searching process. The pruning factor was adjusted to 240.0, the grammar scale to -10, and the insertion penalty to 10.0. These values gave optimum results in both %Correct and %Accuracy. Figure 4.3 shows the results of the %word accuracy and %word recognition.

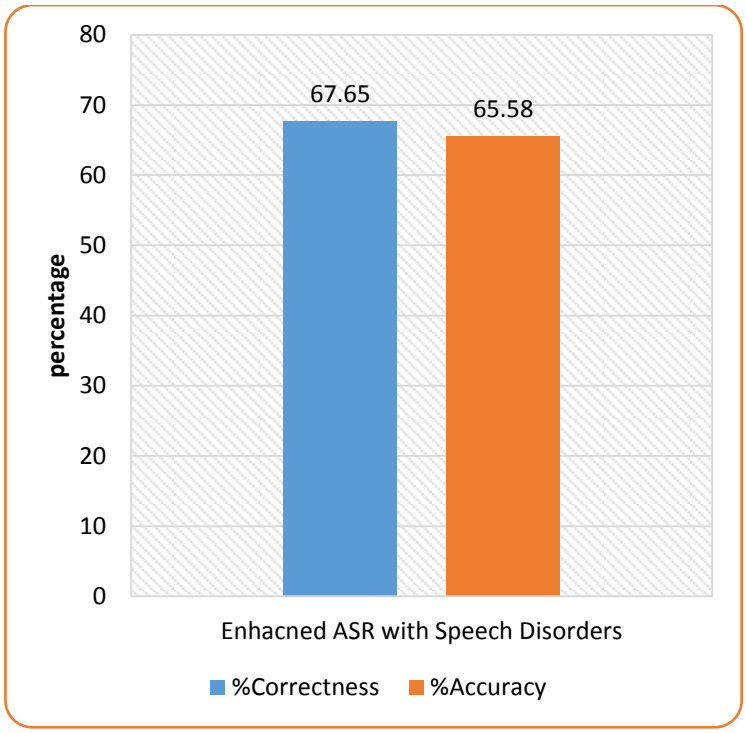

Figure 4-3: Enhanced ASR system tested with disordered speech

As reflected in the results above in Figure 4.3, the new settings of the HDecode resulted with the word recognition accuracy of around 66%. Figure 4.4 compare the results from an ASR baseline and enhanced ASR system tested with disordered speech.

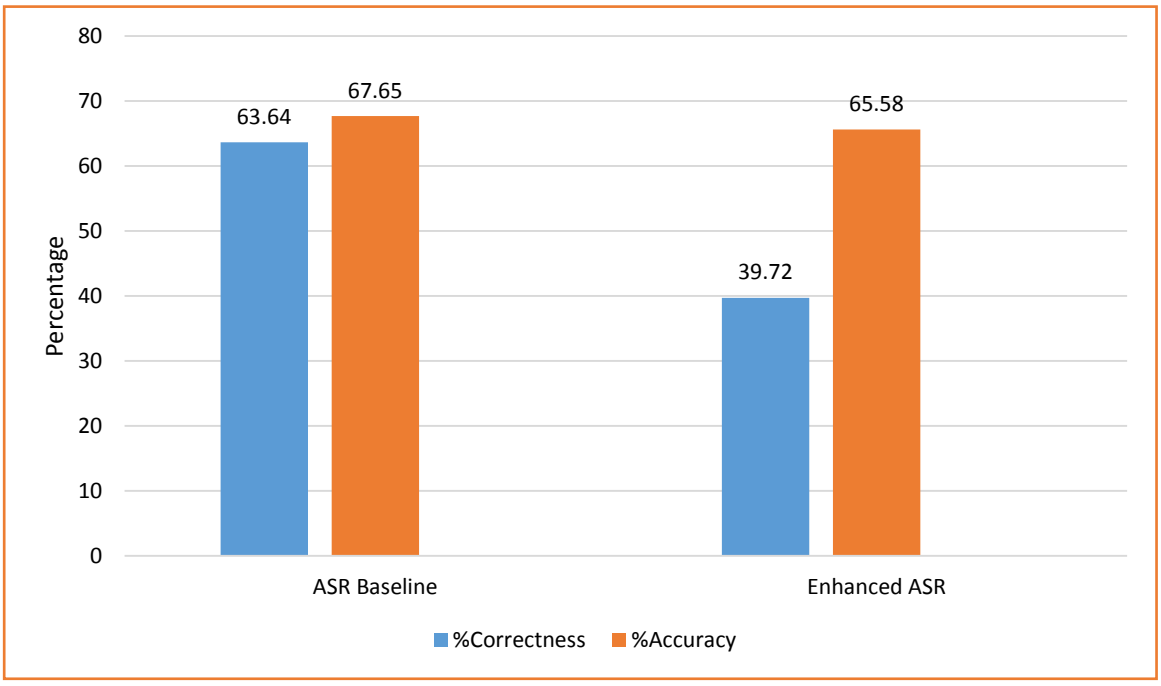

Figure 4-4: ASR Baseline and Enhanced ASR

Comparing these two systems, we can see that the parameters in the HDecode have an impact of the recognition accuracy. Word recognition accuracy was significantly improved by approximately 23%

There are 4 different types of speech disorders that usually occur in stuttering: repetitions, filled pauses, broken words and prolonged sounds or words. As mentioned in the previous chapter, the main focus of this research was on recognition of repetitions. Speech dysfluencies like *uh","um"* and *"eh"* were treated as noise markers, hence, they were not modelled acoustically. In word-level recognition, prolonged sounds and broken words were automatically recognised as words rather individual's phoneme. Therefore they were automatically handle by the HTK-embedded Viterbi search algorithm.

To evaluate the recognition accuracy of repeated words, the word recognition rate (WR) was used (Maier et al., 2010). The measurements were defined as follows:

$$WR\ [\%] = \frac{C}{R} * 100 \qquad (3)$$

Where:

- C is the total number of repetitions correctly recognised as repetitions, and
- R is the number of words in the reference.

Table 4.1 present the summary of WR of repetitions. The values are based on the number of occurrences of repetitions in each speaker.

Table 4-1: Summary of the recognised repetitions

No. of repetitions correctly recognized.	No. of words in the reference.	WR%
63	73	86.3

The original test transcriptions had 73 as the number of repeated words in total. An ASR system was able to recognise 63 repeated of words. The reason why the system did not recognised all the repeated words is a matter of concerned. The

proposed algorithm was applied to remove the repetitions in the MLF. The WR of 86% was achieved. The final output was a normal text without any repetitions of words.

4.3 Conclusion

In this study the main goal is to recognise speech disorders and convert it to equivalent text without disorders. The output of the system is a normal text. The ASR system trained with a normal speech was able to recognise some of the speech with disorders. The proposed refinement approach was applied to remove repetitions in the recognised text. The experiments shows that the proposed algorithm system was able to correctly remove all repetitions in the recognition text.

The next chapter provides a summary of findings and recommendations for future work.

5. SUMMARY, RECOMMENDATIONS, CONCLUSION

5.1 Summary

In this study, an ASR system that can handle both normal and disordered speech has been successfully developed. English normal speech data was used to train the acoustic models. Stuttered speech data was used to test the system. To prepare speech data, features for both data sets were extracted with the TARGETKINDS = MFCC_0_D_A_Z as the energy component. SRILM toolkit and the Lwazi orthographic transcriptions were used to build LMs for training. UCLASS transcriptions were used to build LMs for testing purposes. Pronunciation dictionary Lwazi English dictionary was used to train acoustic models for each entry. However, for the testing process, a new a new pronunciation dictionary called *dict* was created using transcriptions of the test data. The two dictionaries were then merged into one pronunciation dictionary (*baseline.dict*) to accommodate all the missing words in either dictionary. The values of the insertion penalty, grammar scale factor and beam-width pruning threshold were optimally set for decoding. Each test file was recognised and the transcriptions output were stored in a file called MLF. The last step of the recognition phase was to apply the proposed algorithm to remove repeated words in the recognised utterances. The approach showed that the 86% of repeated words were removed from the recognised utterances. The approach showed that the recognised utterance of the system can be used in an application system developed for normal speech.

During the development the major challenge was data preparation. Initially the plan was to use the Northern Sotho speech data that is available at the University of Limpopo Telkom Centre of Excellence for Speech Technology (ULCoE4ST) to train the systems. For testing, the aim was to recruit people with speech disorders and record their speech. It was very difficult to recruit speakers with speech disorders, because there are few people with this type of disorders and also, it is very difficult to find these people because it is not easy to ask people if they are stutters. The other option was to use artificial speech disorders but we wanted to build the system with real speech since this system is going to use by actual people not artificial.

Despite those challenges, 1080 utterances of speech with disorders from six speakers were collected. Unfortunately, the collected speech data could not be used since it was having errors. Some of the audios were not recognised during the training due the poor quality of the recording instrument used. After so many challenges with data collection, it was realised that the system approach does not depend on the language. What is needed is an algorithm to remove repetitions in the recognised text. An English data that is freely available at Lwazi website was used to train the system. To test if the algorithm is working, a stuttered speech that freely available data obtained from the University College of London Archive Stuttering Speech was used. During testing process, it was also a challenge to achieve a good recognition accuracy with disordered speech. The recognition accuracy was much lower when tested with disordered speech. The acoustic model set of the disordered speech were improved by setting the HDecode parameters to optimum values. It was found that there is an improvement in recognition accuracy.

5.2 Recommendations

Speech with disorders is an inherently sparse data domain. By sparse data, we mean a data that have a small or economically disadvantaged user base which are typically ignored by the commercial world (Chan and Rosenfeld, 2012). In order to improve recognition accuracy, it is essential for the system to estimate new more acoustic models using speech with disorders. More meaningful features are needed. Additionally, values of the HDecode parameters can also be used to enhance recognition accuracy of disorders.

5.3 Future Work

Given the success of this research project, we aim to evaluate our system on a live recognition. We also aim to develop a graphic user interface (GUI) for our system and allow potential end-users to evaluate its usability. Our future work includes data collection and evaluation of the performance of the ASR system using acoustic models which are trained by speech samples of disorders. It also includes

expanding this research project further by developing a multilingual speech recognition system for the under-resourced languages South Africa. This would of course require much more time as this would need collection of more domain specific speech data.

6. REFERENCES

Al-Alaoui, M.A., Al-Kanj, L., Azar, J., and Elias Yaacoub, (2008). *Speech Recognition using Artificial NeuralNetworks and Hidden Markov Models*, IEEE MULTIDISCIPLINARY ENGINEERING EDUCATION MAGAZINE, VOL. 3,

Barnard, E., Davel, M. & van Heerden, C. (2009) *ASR corpus design for resource-scarce languages. In: Proc. INTERSPEECH*. pp. 2847-2850.

Badenhorst, J. van Heerden, C. Davel, M. & Barnard, E. (2011). *Collecting and evaluating speech recognition corpora for 11 South African languages*. Language Resources and Evaluation. Vol. 45. pp. 289-309.

Black, A., Lenzo, K. & Pagel, V. (1998). *Issues in building general letter to sound rules.* In Proceedings of the ESCA Workshop on Speech Synthesis. Australia. pp. 77-80.

Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E. & Torres-Carrasquillo, P.A. (2006).*Support vector machines for speaker and language recognition.* Elsevier pp. 210–229

Chan, H.Y. & Rosenfeld, R. (2012). *Discriminative pronunciation learning for speech recognition for resource scarce languages*. In: Proceedings of the 2nd ACM Symposium on Computing for Development. Article No. 12.

Chee L.S., Chia, O. & Sazali, Y. (2009) *Overview of Automatic stuttering Recognition System.* In Proceeding of the International Conference on Man-Machine Systems (ICoMMS).

Christensen, H., Casanueva, I., Cunningham, S., Green, P. & Hain, T. (2014). *Automatic selection of speakers for improved acoustic modelling: Recognition*

of disordered speech with sparse data. In Spoken Language Technology Workshop, SLT'14, Lake Tahoe.

Czyzewski, A., Kaczmarek, A. & Kostek, B. (2003). *Intelligent processing of stuttered speech*. Journal of Intelligent Information Systems, vol. 21, pp. 143-171.

Davel, M.H. & Martirosian, O. (2009). *Pronunciation dictionary development in resource-scarce environments*. In Proceedings of INTERSPEECH. Brighton, UK. pp. 2851-2854.

Ghai, W. & Singh, N. (2012). *Literature Review on Automatic Speech Recognition*. International Journal of Computer Applications (0975 – 8887), Vol 41. No.8. pp. 42-50.

Hollingshead, K. & Heeman, P. (2004). "*Using a Uniform-Weight Grammar to Model Disfluencies in Stuttered Read Speech*: A Pilot Study," Center for Spoken Language Understanding.

Huang, X., Acero, A. & Hon, H. (2001). *Spoken Language Processing: A guide to Theory, Algorithm and System Development*. Prentice Hall, Inc.

Huang, X. & Lee, K.F. (1993). *On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition*. IEEE Trans. On Speech and Audio Processing, pp. 150-157,

Husni, H. & Jamaludin Z., (2010) *Improving ASR performance using context-dependent phoneme models*, vol.12, pp.56-69.

Juang, B.H. & Rabiner, L.R (2004). *Automatic Speech Recognition – A Brief history of the technology development*. Georgia Institute of Technology, Atlanta,

Kitzing, P., Maier A., & Lyberg, A. (2009). *Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders*, Logopedics Phoniatrics Vocology, vol. 34, (2), pp. 91–96.

Maier, A., Haderlein, T., Eyshold, U., Rosanowski, F., Batliner A., Schuster, M. & Nöth E. (2009). *A system for the automatic evaluation of voice and speech disorders* Speech Communication Volume 51, pp. 425-437.

Manaileng, M. & Manamela, M. (2014). *Graphemes and Phonemes as Acoustic Sub-word Units for Continuous Speech Recognition of Under-resourced Languages. Southern Africa Telecommunication* Networks and Applications Conference (SATNAC) 2014, ISBN: 978-0-620-61965-3. Nelson Mandela Bay, Eastern Cape, South Africa. pp. 93-98.

Meraka-Institute. Lwazi ASR corpus (2009). Online: http://www.meraka.org.za/lwazi Accesses on April 2013.

Modipa, T. & Davel, M.H. (2012). *Pronunciation Modelling of Foreign Words for Sepedi ASR.* In 21st Annual Symposium of the Pattern Recognition Association of South Africa, PRASA 2012, Stellenbosch, South Africa, pp. 185-189.

Muhammad G., Mesallam T.A., Malki K.H., Farahat M., Alsulaiman M. & Bukhari M. (2011). *Formant analysis in dysphonic patients and automatic Arabic digit speech recognition.* BioMedical Engineering Vol. 10:41 pp.1-12,

Nöth E., Niemann, H., Haderlein, T., Decher, M., X Eysholdt, V., Rosanowski, F. & Wittenberg, T. (2000). *Automatic Stuttering Recognition using Hidden Markov Models,* vol. 4, pp. 65-68.

O'Brien, S.M. (1993). *Knowledge-based systems in speech recognition: a survey.* International Journal of Man-Machine Studies, Vol. 38, pp. 71-95.

Rabiner, L.R. & Juang, B. H. (2004). *Statistical Methods for the Recognition and Understanding of Speech*. Volume number, 1-35,

Rabiner, L. R. & Juang, B.H. (1993). *Fundamentals of speech recognition,* PTR Prentice Hall Engelewood Cliffs, New Jersey.

Ravikumar, K.M. & Rajagopal R., & Nagaraj H. C. (2009). *An approach for Objective of Stuttered speech Using MFCC Features.* DSP Journal, vol. 9, pp. 19-24.

Ravikumar K.M, &. Reddy, B., Rajagopal, R. and H. Nagaraj, H.C. (2008). *Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies*, in Proceedings of World Academy Science, Engineering and Technology, pp. 270-273.

Ravikumar K.S., Rajagopal, K.R. & Nagaraj, H. (2008) *Development of procedure for the Automatic recognition of Disfuencies in the speech of People Who Stutter.* International Conference on Advanced Computing Technologies, Hyderbad, India, pp. 514-519, .

Reda A. & Khoribi E.L. (2008). *Support Vector Machine Training of HMT Models for Land Cover Image Classification*. ICGST-GVIP, vol.8, pp.7-11.

Ruben R. J. (2000). *Redefining the survival of the fittest: communication disorders in century.* Laryngoscope, vol. 110, no. 2, part 1, pp. 241–245.

Salama E.S., EI-Khoribi R.A. & Shoman M.E., (2014) *Audio-Visual Speech Recognition for people with Speech disorders without Phonemes*. International Journal of Computer Applications, vol. 96, pp. 0975-8887.

Schuster M., Maier A. & Haderlein, T. (2006). *Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition*. International Journal of Paediatric Otorhinolaryngology, vol. 70, no. 10, pp. 1741–1747.

Shunsuke I., Oda K. & Nakayama M., (2011) *Body-conducted speech recognition in speech support system for disorders*, International Journal of Innovative, vol. 7, pp. 4929-4940.

Stolcke, A., Grezl, F., Hwang, M., Lei, X., Morgan, N. & Vergyri, D. (2006). *Cross-domain and cross-lingual portability of acoustic features estimated by multilayer perceptrons*. In Proc. IEEE ICASSP-2006. pp. 321-324.

Stolcke, A. (2002) *SRILM —AN EXTENSIBLE LANGUAGE MODELING TOOLKIT*

Świetlicka, I., Kuniszyk-Jóźkowiak, W. & Smołka E. (2009). *Artificial Neural Networks in the Disabled Speech Analysis*, In *Computer Recognition System 3*. vol. 57, pp. 347-354

Tian-Swee, T., Helbin L., Ariff A. K., Chee-Ming, T. & Salleh, S. H. (2007). *Application of Malay speech technology in Malay Speech Therapy Assistance Tools*.: Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference pp. 330-334.

Van Heerden, C., Barnard, E. & Davel, M. (2009). *Basic speech recognition for spoken dialogues*, In Proc. INTERSPEECH. Brighton, UK. pp. 3003-3006.

Van Nuffelen, G., Middag, C., De Bodt, M. &. Martens, J. P. (2008). *Speech technology-based assessment of phoneme intelligibility in dysarthria*, International Journal of Language and Communication Disorders, vol. 30, pp. 1–15.

Whittaker, E.W.D., & Woodland, P.C., (2001). *Efficient class-based language modelling for very large vocabularies.* In: ICASSP-2001, Salt Lake City, USA, pp. 545–548

Wiśniewski, M., Kuniszyk-Jóźkowiak, W., Smołka, E., & Suszyński, W., (2007). *Automatic Detection of Disorders in a Continuous Speech with the Hidden Markov Models Approach*, in *Computer Recognition Systems 2*, vol. 45, pp. 447-453.

Young, S. et al., (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

Young, V., & Mihailidis, A. (2010*). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review.* Assist Technol. 22, pp.99-112.

Young, S. *A (1996) Review of Large-Vocabulary Continuous-Speech*. Signal Processing Magazine. IEEE. Vol. 13. No. 5

Young, S., (2008), *HMMs and Related Speech Recognition Technologies*. In: Springer Handbook of Speech Processing. Springer-Verlag. Berlin Heidelberg. pp. 539-557.

Young, V., & Mihailidis, A. (2010). *Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review*. Assist Technol. 22(2), pp. 99-112

Yunbin D. et al. (2009). *Disordered Speech Recognition Using Acoustic and sEMG Signals*, Proc. Interspeech, pp. 644-647.