A Comparative Study of Classification Techniques for English to Arabic Speech Recognition

دِراسَة مُقارَنة لتَقنيات التَصنيف للتَعْرُّف عَلىْ تَرجَمَةِ الكَلامِ مِنَ اللُغَةِ الإنجِليزيَةِ إلى اللغَةِ العَربيةِ

Prepared by

Ayoub Abdelrahman Al-Omari

Supervisor

Dr . Hebah H. O. Nasereddin

A Thesis Submitted In Partial Fulfillment of the Requirements of the Master Degree in Computer Information Systems

Department of Computer Information Systems

Faculty of Information Technology

Middle East University

Amman , Jordan

May, 2016

Authorization Statement

I, Ayoub Abdelrahman Al-Omari, authorize the Middle East University to supply a copy of my thesis to libraries, establishments or individuals upon their request.

Name : Ayoub Abdelrahman Al-Omari

Date : 24 / 5 / 2016

Signature : .....Ayoub.....

اقرار تفويض

أنا أيوب العمري، أفوض جامعة الشرق الاوسط للدراسات العليا بتزويد نسخ من رسالتي ورقيا و الكترونيا للمكتبات أو المنظمات أو الهيئات و المؤسسات المعنية بالابحاث والدراسات العلمية عند طلبها.

الاسم : أيوب عبدالرحمن العمري

التاريخ : 2016/5/24

التوقيع : .....Ayoub.....

Examination Committee Decision

This is to certify that the thesis entitled "A Comparative Study of Classification Techniques for English to Arabic Speech Recognition" was successfully defended and approved on 24th May 2016.

Examination Committee Members	Signature
(Head of the Committee) **Dr. Mohammed A. F. Al-Husainy** Associate Professor Middle East University	.....
(Supervisor) **Dr . Hebah H. O. Nasereddin** Associate Professor Middle East University	.....
(External Committee Member) **Prof. Ahmad Ta'ani** Full Professor Department of Computer Science Yarmouk University	....

Acknowledgment

This thesis would not have been possible without the support of many people. Many thanks to my supervisor, Dr. Hebah Nasereddin, who read my numerous. Also thanks to my colleagues, for their help and support. Thanks to the Middle East University for giving me the enhance to be one of its students. Finally, thanks to my parents, and my numerous friends who endured this long process with me, always offering support and love.

Dedication

This Thesis is dedicated to the people who gave me everything and waited from me nothing in return… my parents Abdul-Rahman and Wafa Omari, My beloved wife Rawan Al-Louzy.

To My lovely Sisters Hana, Hanan and Eman, My brother Omar.

To my friend Saddam for his support, to my future kids Omar and Thalia.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Meaning
AC	Acoustic Feature
AM	Acoustic Model
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BN	Bayesian Network
CNN	Convolutional Neural Networks
CVC	Consonant-Vowel-Consonant
DAG	Direct Acyclic Graph
DBN	Dynamic Bayesian Networks
DCT	Discrete Cosine Transformation
DNN	Deep Neural Network
DTW	Dynamic Time wrapping
HMM	Hidden Markov Model
JPD	Joint Probability Distribution
LM	Language Model
LPCC	Linear Predictive Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
MT	Machine Translation
PD	Punctuation Dictionary
PLP	Perceptual Linear Prediction Coefficient
QV	Quantization Vector
RV	Random Variables
SLT	Spoken Language Translation
SVM	Support Vector Machine
WER	Word Error Rate
WRR	Word Recognition Rate

A Comparative Study of Classification Techniques for English to Arabic Speech Recognition

Prepared By

Ayoub Abdelrahman Al-Omari

Supervised By

Dr . Hebah H. O. Nasereddin

ABSTRACT

Speech processing is considered to be one of the most important application area of digital signal processing. Speech recognition and translation systems have consisted into two main systems, the first system represents an ASR system that contains two levels which are level one the feature extraction level As well as, level two the classification technique level using Data Time Wrapping (DTW), Hidden Markov Model (HMM), and Dynamic Bayesian Network (DBN). The second system is the Machine Translation (MT) system that mainly can be achieved by using three approaches which are (A) the statistical-based approach, (B) rule -approach, and (C) hybrid-based approach. In this study, we made a comparative study between classification techniques from ASR point of view, as well as, the translation approaches from MT point of view. The recognition rate was used in the ASR level and the error rate was used to evaluate the accuracy of the translated sentences. Furthermore, we classified the sample text audio files into four categories which were news, conversational, scientific phrases, and control categories.

The empirical findings showed that the DBN achieved the best recognition rate for news category with 79.2% compared with HMM and DTW. However, the HMM

classification technique achieved the highest accuracy in term of recognition rate for conversational with 80.1%, scientific phrases with 86%, and control with 63.8 % recognition rates. In contrast, using DTW in ASR had a negative behavior on the recognition rate for all speech categories. The rule-based model which was represented by IBM Watson cloud achieved high translation accuracy results for the majority of speech categories with 13.93% in conversational, 7.38% in scientific phrases, and 17.91% in control categories. However, by using the statistical-based model – that was represented by Google Translate - in translation the empirical findings showed that for conversational and scientific phrases the error rate was close to rule based with an intangible difference. In contrast, by using the hybrid-based model influenced the error rate in the three ASR classification techniques and for all speech categories which was assigned as a negative effects.

Keywords: Automatic Speech Recognition (ASR), Speech Detection, Speech Translation, Machine Translation (MT).

دِراسَة مُقارَنة لتَقنيات التَصنيف للتَعْرُّف عَلىْ تَرجَمَةِ الكَلامِ مِنَ اللُغَةِ الإنجِليزيَةِ إلى اللغَةِ العَربيةِ

إعداد

أيوب عبدالرحمن العمري

اشراف

د. هبة حسن ناصر الدين

الملخص

تعتبر أنظمة معالجة الكلام واحدة من أهم المجالات التطبيقية في أنظمة معالجة الإشارات الرقمية. وتتألف أنظمة التعرف على الكلام والترجمة إلى نظامين الرئيسية، ويمثل أول نظام من وجهة نظر ال ASR يحتوي على مستويين وهما مستوى استخراج الميزة باستخدام MFCC، من خلال استخدام تقنيات التماثل الثلاث المشهورة باستخدام DTW، و HMM، وDBN. النظام الثاني هو نظام MT والذي أساسا تم تحقيقه من خلال استخدام ثلاثة أساليب التي هي المنهج المعتمد على الإحصائيات والمنهج المعتمد على أساس القواعد، والمنهج القائم على الخلط بينهما والذي سمي بالمنهج الخليط. في هذه الدراسة قدمنا دراسة مقارنة بين تقنيات التصنيف من وجهة نظر ASR، وكذلك الترجمة المعتمدة على المناهج من وجهة نظر MT. ولذلك فقد تم استخدام معدل الاكتشاف وكذلك استخدمنا معدل الخطأ لتقييم دقة الجمل المترجمة. وعلاوة على ذلك، لقد قمنا بتصنيف الملفات الصوتية إلى أربع فئات التي كانت أخبار، التخاطب، والعبارات العلمية.

أظهرت النتائج التجريبية أن استخدام تقنية DBN كأسلوب تصنيف حققت أفضل معدل اكتشاف بنسبة (79.2٪) مقارنة مع HMM وDTW عن فئة الأخبار. ومع ذلك، حققت تقنية تصنيف HMM على أفضل معدل اكتشاف بنسبة (80.1٪) لفئة الاخبار، والعبارات العلمية بنسبة (86٪)، وعبارات السيطرة بنسبة (63.8٪). في المقابل، وذلك باستخدام DTW كأسلوب تصنيف في ASR كان سلوك سلبي على معدل الاكتشاف لجميع فئات الكلام. لقد قمنا بدراسة ثلاثة مترجمات

على الانترنت لإظهار آثار نوع الكلام وتقنية تصنيف ASR على معدل خطأ في الترجمة. ولذلك، فإن النموذج المستند على القاعدة التي تم تمثيله باستخدام سحابة IBM قد حقق أعلى النتائج لغالبية فئات الكلام مع (13.93٪) في التخاطب،(7.38٪) في العبارات العلمية، و(17.91٪) في فئات السيطرة. ومع ذلك، باستخدام نموذج يستند الإحصائية -التي كان يمثلها مترجم جوجل المباشر - ترجمة أظهرت النتائج التجريبية أن نتائج الفئتين التخاطب والمصطلحات العلمية كانت نسبة الخطأ بفرق غير ملموس مقارنة مع استخدام مبدأ القواعد. في المقابل، باستخدام نموذج الترجمة الهجينة فلقد أثرت على معدل الخطأ في تقنيات التصنيف ASR ثلاثة ولجميع فئات الكلام الذي تم تعيينه بوصفه واحد من الآثار السلبية.

الكلمات المفتاحية: اكتشاف الكلام الأتوماتيكية، اكتشاف الكلام، ترجمة الكلام، ترجمة الالة.

CHAPTER ONE

INTRODUCTION

1.1. Overview

In this chapter, the discussion of the Automatic Speech Recognition (ASR) was presented, as well as, the problem statement, research objectives, motivation, and author's contribution were outlined also.

1.2. Automatic Speech Recognition

The most important tool for the interaction between the human being is the speech. Thus, using speech human beings can easily communicate, explain, and investigate their ideas in different fields of life. Hence, human beings would like to interact with computers via speech, rather than using primitive interfaces such as keyboards and pointing devices (Vimala and Radha V., 2012). Therefore, achieving the interaction between the human beings and the computer could be established using Automatic Speech Recognition (ASR) systems. Several computer applications employing speech recognition functions such as electronic dictionaries, customer call centers in communication organizations, the modern generation of automobiles, or the smart house security detecting and authorization processes or more.

The main aim of ASR systems is the transcription of human speech into spoken words. It is a very challenging task because human speech signals are highly variable due to various speaker attributes, different speaking styles, uncertain environmental noises, and so on (Abdel-Hamid, O. Abdel-Hamid, O., Mohamed, A., Jiang, H., Peng, L., Penn, G., and Yu, D., 2014). Thus, the key components of ASR systems namely the Acoustic Feature (AF),

the Language Model (LM), the Pronunciation Dictionary (PD), the Acoustic Model (AM), and the decoder. Figure 1.1 shows the key components of ASR.

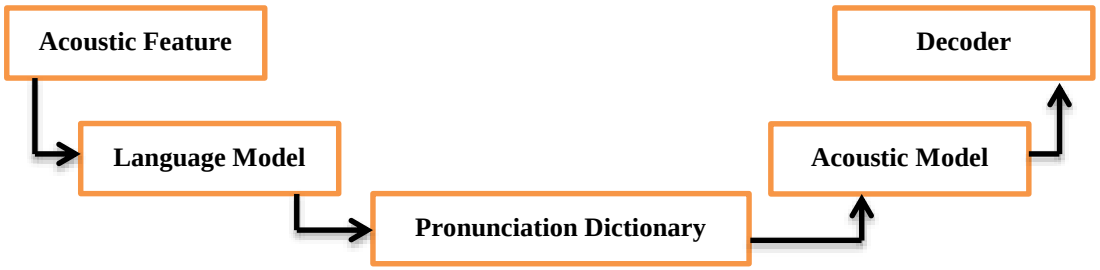


Figure 1.1. The key components of ASR systems (Baker, J. et. al, 2009)

1.3. The anatomy of ASR system

In this section a detailed analysis was presented for the key components of ASR systems.

a. The Acoustic Feature (AF)

The raw data in this level is a raw audio signal that is transmitted from the microphone which needs to be converted into a manageable form in order to make it capable to deal with speech recognition tasks. The input audio signal is converted into a series of sequential frames that are divided based on a specific time interval. Thus, the redundant data is eliminated in this level in order to obtain the representative vector for each frame in the signal.

b. The Language Model (LM)

This component is responsible on describing the combination of words in the target audio signal.

c. Pronunciation Dictionary (PD)

This component is represented as a container of the words and its pronunciation of the source language, as well as, a set phonemes are used for the acoustic models. Furthermore, multiple entries can appear for a word depending on the pronunciation that called homonyms.

d. The Acoustic Model (AM)

This component contains a data describing the acoustic nature of all phonemes that was understood in the system. Thus, the AM usually specific for one language and could be adjusted for a particular language accent. Furthermore, each context dependent phoneme called triphone. One challenge in this level is that the phoneme are context dependent so it important to tend to sound different based on the next and previous.

e. The Decoder

The most important component on the ASR systems and it was represented as the reason behind the ASR system. For each audio frame there is a process of pattern matching. Hence, the decoder evaluates the received feature against all other patterns. The best match can be achieved when more frames are processed or when the language model is considered.

1.4. Feature Extraction

This process is done using a speech signal processor. Thus, it aims to provide a compact encoding of speech wave form. Therefore, the encoding should minimize the loss of information in order to provide a suitable match with the distributed assumptions that was made by the acoustic model. The final results of feature extraction is represented in a feature vector that have specific feature dimension (e.g. 40) that is compared repetitively every specific time interval (e.g. 10 ms) using the overlapping analysis

window. Several techniques found in the literature to employ feature extraction process such as Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), and Perceptual Linear Prediction Coefficient (PLP). Thus, the extracted feature depends on the used feature extraction technique.

a. Mel-Frequency Cepstral Coefficient (MFCC)

This technique is used to extract the feature by calculating MFCC used in speech recognition based on the frequency domain using Mel scale which is based on the human ear scale. Thus, many reference in the literature showed MFCC more an accurate than time domain features. MFCC was represented as an audio feature extraction technique which extracts parameters from a speech similar to one that used human for hearing speech and reducing the importance of the other information. Furthermore, in this technique the speech signal is divided into time frame that contain an arbitrary number of samples, each time frame is windowed with hamming window to eliminate discontinuities at the edges. Consequently, the MFCC calculates the discrete cosine transformation (DCT) of the output from the filter, Mel Bank is a systematical bank that used to classify the coefficient values and eliminate the zero coefficients since it is unreliable.

b. Linear Predicative Cepstral Coefficients (LPCC)

In this technique the concentration is on calculating the power of spectrum of the signal speech analysis filter to remove the redundant data in audio signals. The output that is generated of this technique called residual error. This technique takes into consideration converting the audio signal into quantized bits instead of transferring the entire signal because of the usability of generating the original signal. By using this technique, the speech signal is approximated as a linear combination of previous samples. The obtained LPC coefficient is used to describe the formanant by calculating the

frequencies at which the resonant peaks occur that is called formant frequencies that detect the location of formanant in speech signals.

c. Perceptual Linear Prediction (PLP)

In this technique the irrelevant information of speech and thus improves speech recognition rate. PLP technique is an identical case with PLC but core difference that the spectral characteristics have been translated to match human auditory system. PLP experiments have the following perceptional aspects which are the critical band resolution curves, the equal loudness curve, and the intensity loudness power law which knows as cubic root (Dave, N., 2013).

1.5. The classification methods for ASR system's implementation

Many modern ASR systems recognizing and classifying methods should be built using one of the following methods which are (Cutajar, M., et al., 2013):

- Hidden Markov Model (HMM).
- Dynamic Time wrapping (DTW).
- Dynamic Bayesian Networks (DBN)
- Support Vector Machine (SVM).

In order to cover some ASR sub tasks such as acoustic modeling / language modeling two basic techniques can be used which are (Cutajar, M., et al., 2013):

- Artificial Neural Network (ANN).
- Deep Neural Network (DNN).

From HMM point of view, the reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. Using HMMs representing a complete words can be easily constructed (using the pronunciation Dictionary) from HMMs and word sequence probabilities added and complete network searched for best path corresponding to the optimal word sequence. HMMs are simple

networks that can generate speech (sequences of cepstral vectors) using a number of states for each model and modeling the short-term spectra associated with each state with the mixtures of multivariate Gaussian distributions (the state output transition probabilities and the means, variances and mixture weights that characterize the state output distributions). For each word or phoneme, will have a different output distribution; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM for the separate words and phonemes (Gemmeke, J. et al., 2013).

From DTW point of view, it is an algorithm for measuring the similarity between two sequences which may vary in time or speed. Generally, it is a method that allow computer to find an optimal match between two given sequences (i.e. the sequence are wrapped none directly to match each other) (Vamila C., and Radha, V., 2012).

From DBN point of view, A Bayesian Network (BN) is a way of representing the conditional independence properties of set of Random Variables (RV). Thus, the independences are encoded via missing edges in the graph. in order to clarify the idea DBN and BN is consisting of an indefinite number of frames contains two variables which are from state and the observation, and two sedges (state to the observation, from state in the previous frame to the current state) (Livescu, K, Bilmes, J., and Glass, J., 2003).

1.6. ASR system types

Several types of classifications were considered to identify ASR systems. Thus, ASR systems are identified based on the number of words processed per time. Therefore, Benzeguiba, M. et al. (2007) classified the ASR systems based on the speech nature into four categories namely the ASR for spelled speech that contain pauses between letters or phonemes, the ASR for isolated speech that contains a pauses between words, the ASR for spontaneous speech that contain human to human dialog, and the ASR for highly

conversational speech that contains a speech for a meeting discussion of several people (Benzeguiba, M. et al., 2007).

Furthermore, the ASR systems are identified based on the number of vocabularies (i.e. lexicon or the decoder size). Therefore, ASR systems were classified based on this criterion to ASR for the small size which contains vocabularies up to thousand words, ASR for the medium size which can contain up to 100,000 words, and ASR for large size that contain more than 1000k words.

1.7. Machine Translation (MT) approaches

Several MT approaches were found in the literature. In this section we will provide a general overview about the popular and commonly used MT approaches. Figure 1.2 shows the MT approaches.

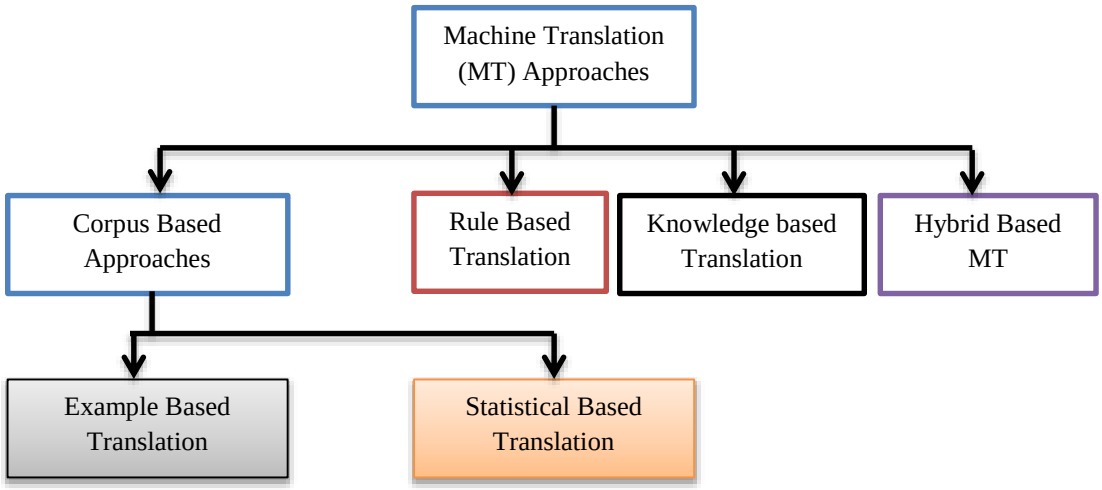


Figure 1.2. Machine Translation Approaches (Mamta, A. and Wala, T., 2015)

a. Rule based translation approach

Actually, this approach relies on the linguistic rules and dictionaries. Thus, the analysis of the source text, the converting between two languages, and the synthesis of the target text are based on the linguistic modules. However, the rule based system requires syntax analysis, semantic analysis, syntax generation, and semantic generation. The analysis in this approach produces a complete parsing of the source language sentence. The final generation of this approach should have lexical transfer, mapping and agreement (Mamta, A. and Wala, T., 2015)

b. Knowledge based translation approach

In this approach the basic idea is to make heavy emphasis on the functionality in order to preserve complete understanding of the source text before the process of translation. This approach assume that an interpretation engine can achieve successful translation into several languages. Thus, the knowledge based approach should supported by world knowledge and linguistic semantic about the meaning of words and their combinations in the target language (Antony, P. J., 2013).

d. Corpus based translation approach

This approach usually divided into two sub approaches which are the statistical based approach, and example based approach. In case of statistical based approach, the translation is generated based on the statistical models. Hence, the parameters of the statistical model is derived by the analysis of bilingual text corpora. Thus, this sub approach brings some benefits such as the linguistic knowledge does not require. In contrast, the challenge in using this approach is the occurrence of massive parallel corpus. On the other hand, the example based translation approach is a knowledge based approach

and it rely on the taxonomy of knowledge and contains an inference engine (Mamta, A. and Wala, T., 2015).

e. Hybrid base MT

This approach usually build due to the weaknesses of two approaches and their possibility to be integrated (i.e. Rule based approach and Statistical MT). This approach contains three basic components which are (1) the identification of source language by observing chunks (i.e. words, phrases, and synonyms), (2) transformation of chunk into the target language, and (3) the generation of the translated language. Hence, the basic workflow of this approach is to make preprocessing and post processing operations to achieve the integration between the combinations (Syahrina, A., 2011).

1.8. Problem Statement

Automatic Speech Recognition (ASR) systems provide an efficient way to extract the spoken text from speech signals by implementing several feature extraction approaches, as well as, employing different types of classification methods. Finding the best feature extraction approach and classification method – with regard to the translated speech- that suits Machine Translation (MT) systems is a challenge. Different approaches were referenced to implement robust MT systems that were varied from statistical approaches to rule based approaches; this represent a challenge in selecting the needed MT system's approach. The main focus in this study was to find the ASR classification technique that suits MT approach that achieve the most accurate translation of a specific type of speech. This study concentrated on finding the effects of classification methods and MT system's approaches on the translated speech. This study took into consideration the Word Error Rate (WER) and Word Recognition Rate (WRR) as an evaluation metrics. Problem will be accomplished by answering the following research questions:

1. What is the suitable ASR classification method approach with regard to speech translation accuracy results?
2. What is the suitable MT based models with regard to speech translation accuracy results?
3. What are the suitable ASR classification technique and MT based model for specific type of speech?

1.9. Objectives

This research aims to find the most accurate translation of a specific speech based on the ASR classification technique as well as the translation model in MT systems. To achieve this aim we proposed to run a set of experiments that covers the several implementations in ASR systems by taking into consideration MFCC as a feature extraction approach with HMM, DTW, or DBN classification techniques in the ASR system, as well as, taking into consideration three MT approaches statistical based translation, rule based translation, and hybrid MT approaches. Performance – with regard to Word Error Rate (WER) and Word Recognition Rate (WRR) metrics - test was calculated for four different environments in term of news, conversational, scientific phrases, and control categories. This study proposed to eliminate the noise factor using endpoint detection algorithm on the proposed experiment. The comparison of accuracy results were conducted to extract the main research variables influence on the target translated text.

1.10. Motivation

The use of implementation of ASR system is rising. Many modern applications adding speech recognition functionality in order to ease the interaction between the human beings and the computer based systems such as cars, telephone application and so

on. For the last ten years, the research in the field of speech recognition concentrated on finding the suitable implementation that suits the target application due to the diversity in ASR implementation of classification techniques. Hence, these classification techniques had its own effect on the accuracy of the detected speech, as well as, each MT model had its own influences on the translated texts. Thus, this motivates me to explain the differences of accuracy metric in real time MT based on finding the effects of ASR classification techniques on the MT approaches in term of the accuracy of the translated text. Therefore, these challenges motivates me to find the suitable implementation of the integration between ASR classification techniques with the MT models to deal with several speech types.

1.11. Contribution

The main contribution at this study found in the following tasks:

- Measuring the accuracy of the translated speech text in Spoken Language Translation (SLT) systems.
- Measuring the recognition rate of the results of the ASR classification technique in each speech category to analyze the way that classification techniques were interacting to extract speech words.
- Measuring the error rate in the translation results to analyze the effects of MT models on a specific type of speech.
- Explaining the factors that affect the selection criteria of MT systems approaches in term of accuracy of speech translation from English to Arabic.

CHAPTER TWO

Literature Review and the Related Research Works

2.1. Overview

This chapter contains an investigation on the existing work that is related to this research and it represents a context of this research.

2.2. Automatic Speech Recognition (ASR)

Watanbe and LeRoux, (2014) studied the use of black box optimization, their hypothesis stand on proposing ASR system that automatically tune without any prior knowledge. Thus, they considered ASR system as a function that contains tuning parameters as input, and the word accuracy as an output. Therefore, they used two probabilistic black box techniques which were covariance mean adaption Bayesian optimization process. They conducted multiple experiments to extract the effects of ASR construction on tuning initialization and task variations. There results showed that even using bad tuning initialization, the ASR system performance was converged to the obtained by human experts. Consequently, we used Watanbe and LeRoux way in finding the end points of sound audio files in order to detect the number of words in each sentence was recorded. In contrast, the difference between their study and ours, that their study was focused on designing a system without training phase in order to detect textual data form speech audio files but in our study for the three classification techniques we used two phases as a black box which were training and testing phases.

Giannoulis, and Potaminos (2012) proposed a hierarchical classification system that was based on the discriminative power of the trained sub systems. They designed each subsystem to make feature selection in two stages algorithm. Their system took into

consideration the advantage of well-established method of gender dependent systems to achieve better results. The results of their system showed that the overall classification accuracy was of 85.18%. The accuracy enhancement of their system achieved due to the selection of glottal-flow and AM-FM features. Actually, we applied their way in calculating the accuracy using WER of the textual data in speech audio files. They took into consideration several feature selection approaches while our study focused on using MFCC as a default feature selection approach since their empirical findings showed that MFCC achieved low WER results.

Yadav and Mukhedkar (2013) compared the performance classification methods in speech recognition systems. Their focus was on capturing Hidden Markov Model (HMM) and Vector Quantization (QV) in Artificial Neural Network (ANN) classification method. Their results showed that the implemented HMM with speaker dependent caused low percentage of performance. Basically, the difference between their study and ours that their study focused on finding and recognizing the speaker information from speaker's voice. They measured the WRR of results after using ANN as classification method.

Ng, R. et al. (2015) presented a method to rescore ASR results to improve translation quality. They trained a translation quality estimation model for English to French speech to text translation systems. They observed an improvements in translation performance in cumulative ASR rescoring. Their experimental results found that rescoring ASR output was more efficient when the ASR sentence had low confidence. They calculated the confidence by finding the mean of word frequency for each rescore level. The performance was computed using baseline and Oracle performance in MT. Actually, their study focused on defining sentence confident in order to rescore textual

data results of MT system. The difference between their and our study that their study focused on comparing the knowledge-based translation system with statistical-based translation system in term of the target sentence confident.

Rajnoha and Pollak (2011) proposed an ASR system that can deal with various noisy environmental conditions. Therefore, they selected two feature extraction approaches which were Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Predictions (PLP) coefficients. They conducted the experiments in real noisy environment using AURORA database. Their experimental results showed that MFIP and BFCC based recognition had a positive effect on speech dynamics for LP feature extraction approach. Furthermore, they found out that enhancing noise reduction techniques reduced the error rate by 40% using AURORA 3. Consequently, their study showed that using L-P feature extraction approach within ASR had a positive recognition performance in noisy environment. From their study we took into consideration their technique for noise removal algorithm. Their study focused on comparing the feature extraction techniques MFCC and PLP in noisy and unnoisy environments and they used to eliminate the noise by adding silence periods in high noise rates. Our study took into consideration the measuring tools of MFCC feature extraction technique because it is the only one technique in our study.

Abdel-Hamid et al. (2014) explained a solution for error rate reduction by using Convolutional Neural Networks (CNN). They proposed a limited Weighted Sharing schema in order to represent speech feature. Thus, they used special structure such as local connectivity and weight sharing in order to exhibit some degree of invariance to small shift of speech features. Consequently, their experimental results showed that CNN

reduced the error rate by 6-10% compared with the feature speech extraction approaches in phone recognition and voice large vocabulary speech recognition task. The difference between their study and ours that their study compared the CNN and feature speech extraction as classification technique in order to measure the error rate of the results. As well as, their study took into consideration the system response time after each speech recognition transaction.

Alsuliaman et al. (2011) developed a rich database for Arabic language. Thus, they showed the richness of the Arabic database in term of text stored, environments, microphone types, number of recognition sessions, the recoding system, and the transmission channel. Hence, they developed large database that was contained of a large number of male and female speakers, variety of text materials. The transmission channel for their Arabic database was assigned to be in three levels two of them based on the type of microphone and the other based on mobile phone. Actually, we used their process in recording audio files in our training phase. The difference with our study that they were working to create Arabic speech audio files while in ours we made a training phase for an English sentences.

Alotaibi and Hussein (2010) focused on investigating the analysis of vowel in modern standard Arabic dialects. They used Consonant-Vowel-Consonant (CVC) utterances to find the similarities and differences between the Arabic vowels. Hence, they employ Hidden Markov Model (HMM) as a recognizer to clarify the vowels, As well as, they extracted the performance of the recognition in term of phonetic feature of vowels to compute the differences and similarities metrics. Consequently, they analyzed the time and frequency domains for classification purposes. We used their technique in our study

in case of computing the similarity in our matching process after comparing the utterance of the speaker with the stored utterance of the same speech in DB. The difference that we added the log energy algorithm after matching process in case of using HMM as classification technique in order to define the short energy of speech signal to detect features correctly. Furthermore, their technique could not be capable to be implemented with continues speech files.

Mon and Tun (2015) discussed an approach to extract features by using MFCC from the speech signals of isolated spoken words. As well as, they used HMM method in order to train and test the audio files to get the recognized spoken word. They created speech database using MATLAB. The original speech signals in their experiments were preprocessed and these speech samples were extracted to the feature vectors which were used as the observation sequences of the Hidden Markov Model (HMM) recognizer. The feature vectors are analyzed in the HMM depending on the number of states. From their simulation results, the average recognition rate of 87.6% achieved by the number of states (N=5) was better accuracy than any other states. But, if the number of states was too large, there were no enough observations per state to train the model. So they conclude that, this may degrade the performance of the system. Thus, the choice of the number of states in the HMM also was playing an important case in recognition. The difference between their study and ours that in their work the experiments were conducted on an isolated words than continuous speech in case of our study. As well as, their study focused on finding the feature vector for each state in the one word and they measured the similarity and dissimilarity from state to another state. Furthermore, in our work, the performance of the system was more accurate and reliable by using end point detection algorithm in preprocessing stage.

Mishra et al. (2016) discussed the process of automatically generating subtitles for videos using Google's free API's. They proposed three step model to realize their process. The first stage was extracting the audio to be transcribed from the video file and then converting it into a format compatible with their second stage. Their second stage used a speech recognition engine to convert the information from audio format to text format. Their third and final stage was the encoding of the generated text into a subtitle format which include adding time frames and necessary pauses and punctuation marks. Their hypothesis was based on using DBN as classification technique. The difference between their study and ours was in the type of classification techniques used and the type of MT system. Therefore, by signing the utterance words with the time of occurrence in the audio file in order to store them speech database. In contrast, their implementation showed weakness in finding the correct speech signal with the suitable words.

Jouvet and Vinusea (2012) investigated class based speech recognition to find the comparative of selection of the training samples for each class on the final speech recognition performance. They presented standard automatic classification procedure to build 2, 4, 8, and 16 classes. Their experimental results showed that by increasing the number of classes there were fewer and fewer training data in each class for adapting the acoustic model parameters. Thus, the results showed unreliable parameter with speech recognition decreasing. Therefore, investigation concentrated on finding the classification margin in the selection of training data. The difference between their study and ours that their speech recognition technique was capable to classify each word into a standard classification in order to be retrieved with their classification number (i.e. speech type in DB) in contrast, our matching criteria took into consideration feature vector, and

Mel Frequency, to find the similarity even if the speech in training set was not identical to the speech in testing set.

2.3. Theoretical background and related works

In this section, the author provide a theoretical background about using the ASR classification techniques. Therefore, DTW, HMM, and DBN techniques were discussed in details.

A. Dynamic Time Wrapping (DTW)

Any two time series can be varied in time and speed which is called wrapping points. Thus, data time wrapping technique is one of the most used feature matching (i.e. classification) techniques. Consequently, this technique is used to find the optimal alignment between two time series, as well as, measuring the similarity between those time series (Zhang et al., 2013).

The DTW is employing linear time wrapping by comparing signals of two time series based on linear mapping of the two temporal dimensions (Chapaneri, 2012). Thus, DTW allow non-linear alignment of one signal to another by minimizing the distance between two signals. Therefore, this wrapping can be used to extract figure recognition based on the similarity and dissimilarity between those signals. From speech signals point of view, the duration of each spoken word or digit can vary but the overall speech waveform are similar for same word or digit. Therefore, by applying the DTW technique the corresponding regions between the two time series can be extracted easily to be used in matching processes (Muda et al., 2010).

In more details, Chapaneri, S. measured the optimal wrap path for a given two time series A and B where the length of A is |A| and the length of B is |B|. Thus, A = A_1, A_2, A_3… $A_{|A|}$, and B = A = B_1, B_2, B_3… $B_{|B|}$, a wrap can be constructed W = W_1, W_2, W_3 … W_k. Where k is the length of wrap path for a k^{th} element that is max (|A|, |B|) ≤ k ≤ |A| + |B|, and W_k = (i, j) where *i* is the length of time series A, and *j* is the length of time series B. Consequently, the optimal wrap path that represents the minimum distance between two time series can be calculated using (equation 1) (Chapaneri, 2012).

$$Dist\left(w\right) = \sum_{k=1}^{|k|} \frac{Dist\left(w_{ki}, w_{kj}\right)}{(|A| + |B|)} \quad (1)$$

In our research, the two time series corresponds to the two number of coefficient features from MFCC phase. Thus, each one of these time series was represented as a vector from different speech signal with two dimensional cost matrix for each feature vector A_i and B_j. The spoken feature vector was compared to template feature vector using DTW (i.e. in case of using DTW classification technique) and the one of the minimum distance is chosen as a recognition output.

B. Hidden Markov Model (HMM) technique

The core idea in using HMM for speech recognition applications is to create a stochastic models from known utterances and compares it with the unknown utterances was generated by speaker. An HMM *M* is defined by a set of states *N* that have *K* observation symbols as well as, three possibility metrics for each state which are in (equation 2) (Ghahramani, 2001)..

$$M = \{\Pi, A, B\} \quad (2)$$

Where:

- ∏: initial state probability.
- A: $a_{t,j}$ state transition probability.
- B: $b_{t,\,j,\,k}$ symbol emission probabilities.

For each HMM system, it could be use three different types of topologies to employ Markov chain which are ergodic model, general left to right model, and linear model. Figure 2.1 illustrates HMM topologies for a system with four states. Consequently, each state has its own probability which leads to compute the probability for an occurrence of state in a given situation of another state using Bayesian rule.

In this context, for any system employs HMM technique three basic algorithms which are classification, training, and evaluation algorithms. In classification algorithm, the recognition process is enabled for any unknown utterance by identifying the unknown observations sequence via choosing the most likely class to have produced the observation sequence. In training algorithm, the model is responsible to store data collected for a specific language (i.e. in our research the language was the English language). In the evaluation algorithm, the probability of an observation sequence is computed for matching processes.

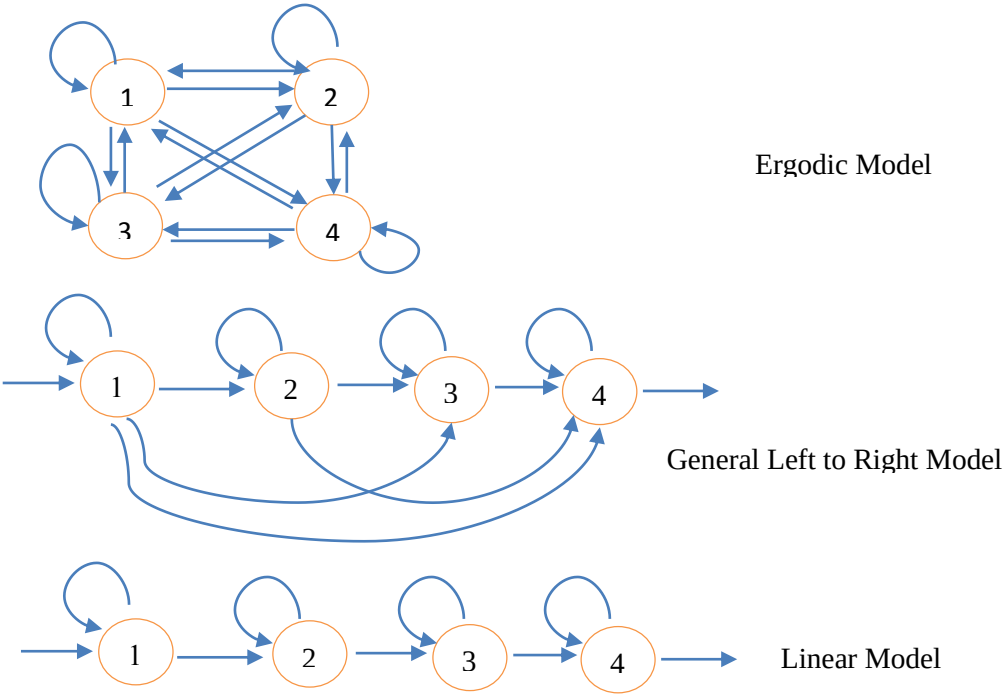


Figure 2.1. HMM three states topologies for a system with four states 1, 2, 3, and 4 (Paul, 1990)

The classification algorithm was employed for a given observations O = O_1, O_2, O_3 … O_T. A chosen class was computed using (equation 3) (Paul, 1990).

$$\text{Chosen\_Class} = \arg MAX \left[P\left(M_{class} \,|O \right) \right] \tag{3}$$

Therefore, by applying Bayesian rule to find $\left[P\left(M_{class} \,|O \right) \right]$ the probability was computed using (equation 4) (Paul, 1990).

$$P\left(M_{class} \,|O \right) = \frac{P(\mathrm{O}|M_{class})\,\mathrm{P}(\mathrm{M}_{class})}{P(O)} \tag{4}$$

In this research, we trained the proposed system by several records. Thus, each record had a several tokens for each word in the vocabulary in order to process them with a front end to create the observation sequences based on each token. Furthermore, each training data was identified by word label (i.e. word spelling). In the recognition process, a probability of each word was calculated in order to match the occurrence of specific

word with another one in the vocabulary table. In contrast, for an unknown observation sequence the same processes were implemented in order to pass them to HMM.

C. Dynamic Bayesian Networks (DBN) technique

Several research works described DBN technique as the general and flexible model because of its capability in representing complex temporal stochastic processes (Franklen et al., 2007). Thus, this technique also called dynamic probabilistic networks. Furthermore, DBN technique include directed edges pointing in the direction of time that provide a computation of Joint Probability Distribution (JPD) among random variables. In contrast to HMM, a DBN technique allow each speech frame to be associated with an arbitrary set of random variable (Garg and Rehg, 2011).

The Bayesian network is defined by a graphical model structure *M* and a family of conditional distribution *F* and their parameters *O*. The model structure *M* consists of a set of nodes *N* and a set of direct edges *E* connecting the nodes which results a Direct Acyclic Graph (DAG). Consequently, the nodes represents the random variable in the network as well as, the edges encodes a set of conditional dependencies. Therefore, in Bayesian network the direction of arrow is important between nodes. For instance if the arrow direction from node 1 to node 2 that means that node 1 influence node 2.

In this context, Stephenson, T. et al. concluded the Bayesian network for a given set of random variables that were denoted by $X=\{X_1, X_2 \ldots, X_n\}$ that correspond to the nodes V in the network. And the values of corresponding variables were denoted by $x = \{x_1, x_2 \ldots, x_n\}$. Hence, the joint distribution over a random variable x as in (equation 5) (Stephenson et al., 2000).

$$P\left(x_1, x_2 ..., x_n\right) = \prod_{i=1}^{n} P\left(x_i Pa\left(X_i\right)\right) \quad (5)$$

A DBN technique is a temporal extension of Bayesian network technique. Figure 3.10 illustrates the DBN model for recognition an isolated word (Arab) that was pronounced as (/'æ/ - /r/ - /ə/ - /b/). Therefore, we employed the following calculations to find the joint distribution of a finite length time series for speech signals. Thus, let X[t] = {X_1[t], X_2[t] …, X_n [t]} to denote to the random variable in X at a time t ∈ {1, 2, 3…}. For all t > 1 and for all values of X [1], X [2] …, X [t] the joint distribution was computed based on (equation 6) (Stephenson et al., 2000).

$$P\left(\mathrm{X}[1], \mathrm{X}[2] ..., \mathrm{X}[t]\right) = P\left(X\ [1]\right) \prod_{t=2}^{T} P\left(x\ [t]\, x\ [t-1]\right) \quad (6)$$

In figure 2.2 the gray vertices (articulators) are observed in training (i.e. when available but not in normal recognition. Hence, the acoustic and final position as well as the transition variables were always observed.

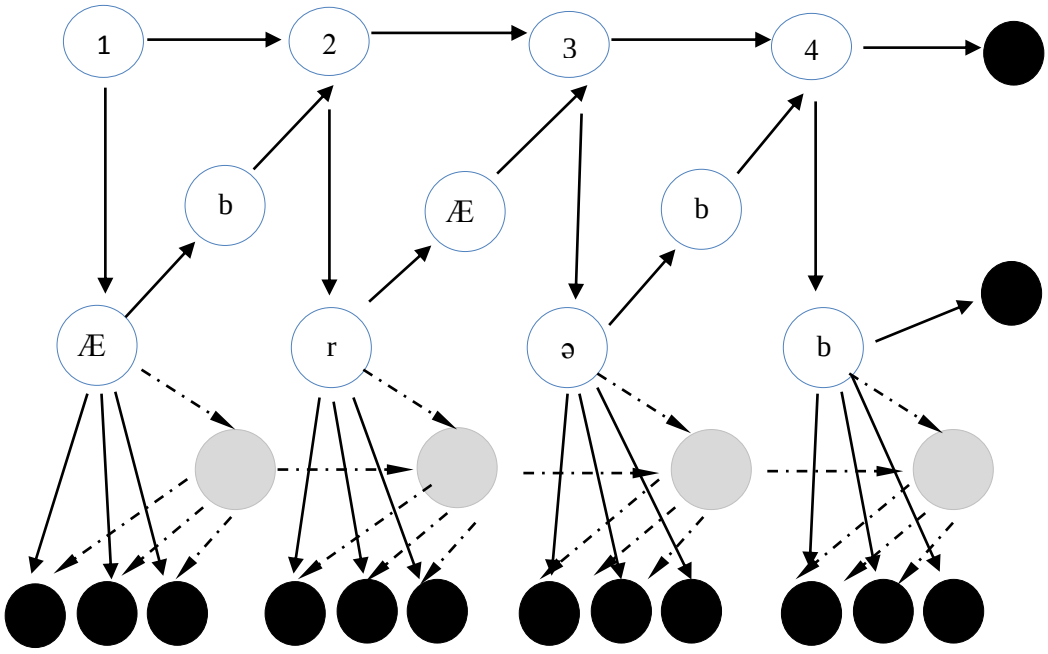


Figure 2.2. a dynamic Bayesian network for isolated word (Arab) recognition covering five time steps

D. Feature Extraction phase using MFCC

By using the technique in feature extraction process a fingerprint is created of sound files (Kurzekar et. al, 2014). Hence, MFCC technique is capable to capture the important properties of audio signals in term of time and frequency (Hasan et. al, 2004). Therefore, to employ MFCC feature extraction technique several steps should be implemented. Hence, MFCC can be implemented using six primary steps which are preprocessing, framing, hamming windowing, Fast Fourier Transform (FFT), Mel bank filtering, and Discrete Cosine Transformation (DCT) steps.

In preprocessing step, the speech input was recorded at sampling rate of 22050 Hz in order to minimize the effects of aliasing because of the conversion from analogue to digital form. Furthermore, in this step the energy of speech signal was increased in order to emphasize a higher frequencies. Hence, the output signal from this step was calculated using equation (7) (Kurzekar et. al, 2014).

$$Y\left[n\right] = X\left[n\right] - 0.95 * X\left[n-1\right] \quad (7)$$

Where:

- Y [n]: The output signal.
- X [n]: The original signal.
- N : The number of sample frames.

In framing step, the speech signal was segmented into small frames (n) of the length which was varied between 20 to 40 msec in order to pass these frames to the hamming windowing step. Consequently, hamming windowing was responsible to create a window shape by considering the next block of feature extraction processing chain as well as integrating all the closest frequency lines. Thus, hamming windows was computed based on equation (8) and (equation 9) (Kurzekar et. al, 2014).

$$Y\left[n\right] = X\left[n\right] * W\left[n\right] \quad (8)$$

$$W\left[n\right] = 0.54 - 0.46 * cos\left(\frac{2\pi n}{N-1}\right) \quad (9)$$

Where:

- W [n] : Hamming window
- N: number of samples in frame.
- Y[n] : output signal
- X[n]: input signal

In FFT step the frame was converted of N samples from time domain into frequency domain to preserve the convolution of glottal pulse and vocal tract impulse response in the time domain. Therefore, the computation in this step was conducted based on equation (10) (Kurzekar et. al, 2014).

$$Y\left[w\right] = \text{FFT}(h\left[t\right] * x\left[t\right] \quad (10)$$

Where:

- h [t]: vocal tract impulse.
- x [t]: glottal pulse.
- Y[w]: Fourier transform of Y[t]

Based on the results of FFT step the spectrum frequencies were very wide as well as, the voice signal does not follow the linear scale. Therefore, the Mel filter bank was used to ease the conversion to get a Mel frequency signal that is appropriate for human hearing and perception. The Mel frequency was computed in this step based on (equation 11) (Kurzekar et. al, 2014).

$$F\left(Mel\right) = \left[2595 * \log_{10}\left[\frac{1+f}{700}\right]\right] \quad (11)$$

Where:

- F (Mel): Mel frequency.
- F : a specific signal frequency

CHAPTER THREE

The Proposed Method and Experiments Design

3.1. Overview

This chapter discusses a detailed description of the proposed method as well as, discuss the proposed experiment's design, and finally define the evaluation criteria used for each research parameter.

3.2. The proposed method architecture

The main theme of this research is to find a suitable ASR system that suits MT system for real time speech translation from English to Arabic. Thus, to meet this aim we ran several experiments to cover several environments and techniques to calculate the accuracy of the recognized and translated sentences in each combination. Therefore, we examined nine environmental combinations to cover feature extraction using MFCC with classification approaches (i.e. HMM, DTW, and DBN), and MT translation approaches (knowledge based approach, rule based approach, and hybrid based approach).

Furthermore, for each combination we applied hundred speech audio files in two clusters which were (i) training cluster with fifty sentences and 297 words for each sentence and (ii) testing cluster with fifty sentences -that covered four speech type categories (i.e. 14 sentences in news category, 15 sentences in conversational category, 10 sentences in scientific phrases category, and 11 sentences in control category). Hence, all speech audio files were recorded on the same microphone type, storage machine, and in the same place to guarantee an identical situation for all speech audio files recording operation. In this context, the noise factor was eliminated in speech audio files by adding

artificial silence between words in order to identify the boundaries of speech. The number of experiments were examined in this study was for training cluster (747) experiments (i.e. 1 type of feature extraction, 3 types of classification techniques, (3) types of MT approaches, (50) sentences were applied, and (297) words) as well as the number of experiments in testing phase was 450 experiments. Therefore, totally the number of experiments in this study was 1197 experiments. Figure 3.1 illustrates the main environmental combination processes examined in this research.

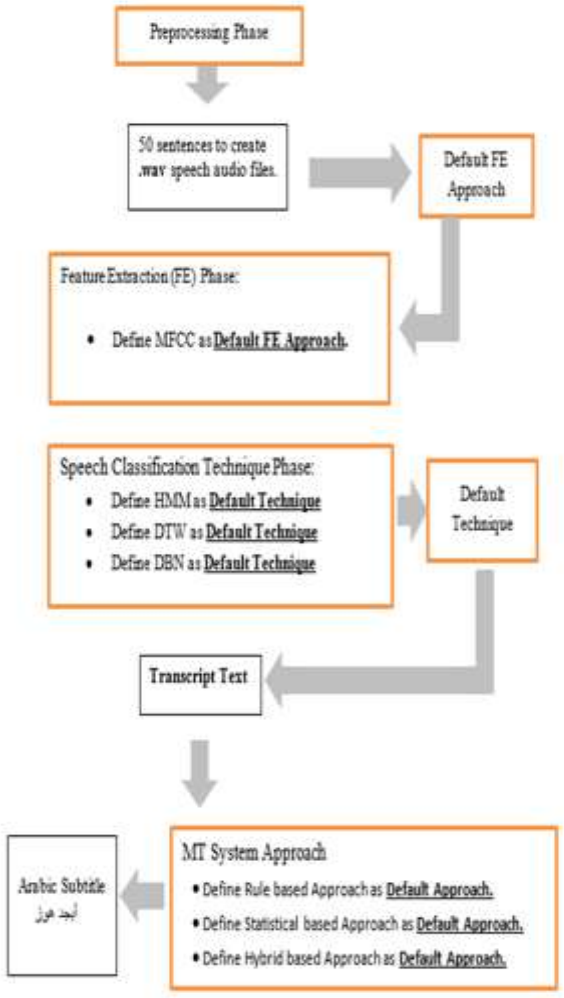


Figure 3.1. The main phases were examined in this research to identify the accuracy and performance of real time speech recognition and translation from Arabic to English

3.2.1. Preprocessing phase

The main goal of this phase is to get the speech signal of each word had been spoken. Thus, the main functions were processed in this phase started from recording the proposed sentences, segmenting the speech audio file by detecting endpoints as well as the silence regions from the utterance, and noise filtering. Thus, the recording process was achieved by recording all speech audio files using the same microphone device and in the same office. Figure 3.2 shows a side of recording process for the data collection of speech audio files.

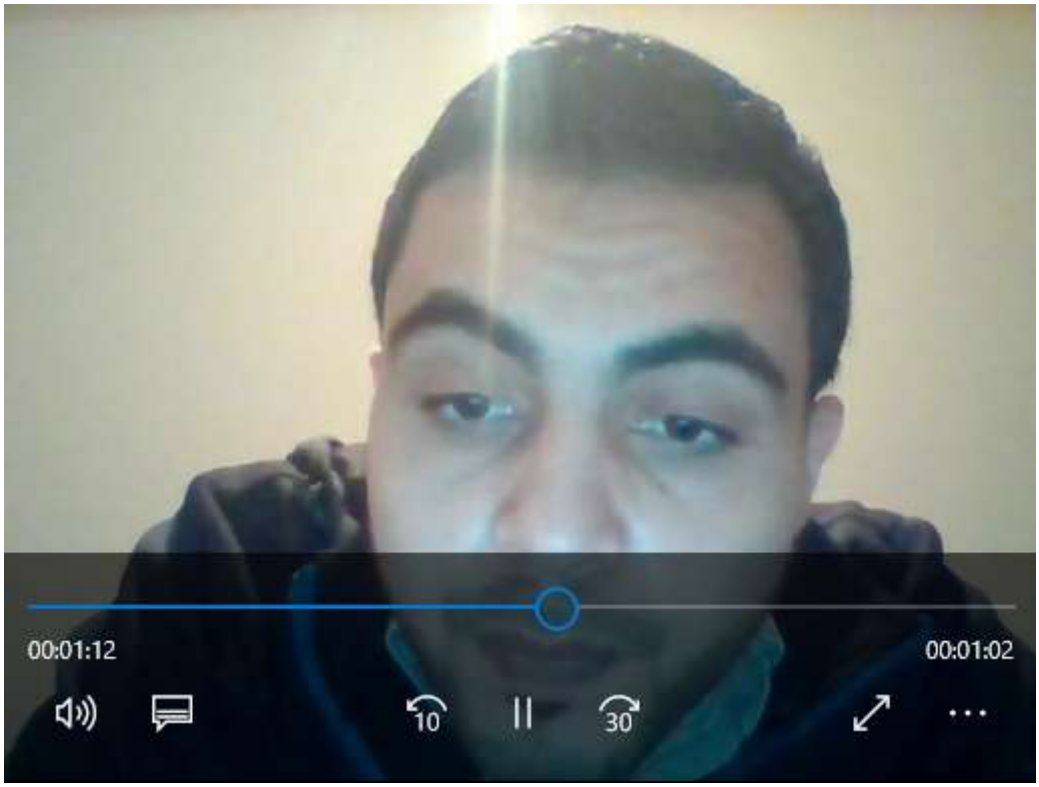


Figure 3.2. side from recording session to produce speech audio files that were used in this study.

In this context, end point and silence regions detection process was conducted using signal log energy algorithm (Rabiner and Schafer, 2014) that is based on analyzing speech frames. Hence, by computing short time log energy and short time zero crossing rate. Thus, Rabiner, L, and Schafer, R. examined the speech signals by computing the mean

(μ) and standard deviation (δ) for the first 1600 sample rate and set them as the initial values because the speaker takes some time to read when recording starts. Their algorithm computes and uses the short-time log energy and the short-time zero crossing rate parameters. The location of the isolated speech utterance within the audio file are determined by using a set of log energy and zero crossing rate thresholds which identify regions of voiced and/or unvoiced speech based on the short-time parameters exceeding a specified set of log energy and zero crossing rate thresholds for a specified number of frames (Rabiner and Schafer, 2014). Figure 3.3 illustrates the end point detection process using log energy algorithm.

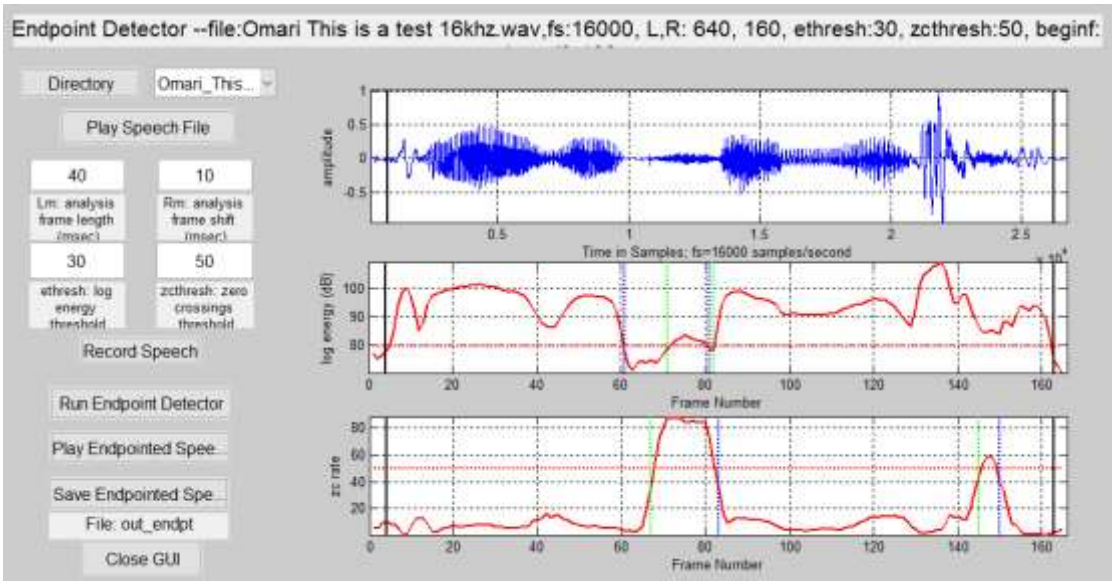


Figure 3.3. detecting the end point of the recorded sentence '*This is a test by Ayoub AL-Omari*' using log energy algorithm GUI.

In Figure 3.3 the speech audio file duration was approximately seven seconds and it contained 163 frames and each frame size was equal (40 ms). The energy threshold that should be exceeded to capture a voice was 30, as well as, the zero crossing threshold was

50. The speed of sample frames was 16000 sample frame per second. The results of the preprocessing phase was summarized as the following:

- Recording speech audio file.
- Detect voiced and unvoiced regions (i.e. find the frame number of voiced regions and the frame number of unvoiced regions)

3.2.2. Feature Extraction phase

Basically, the experiments in this phase were conducted based on the related work were mentioned in the theoretical background in chapter two. Feature extraction process in this study was conducted using MFCC for all speech audio files. Figure 3.4 shows the process of feature extraction using MFCC flowchart.

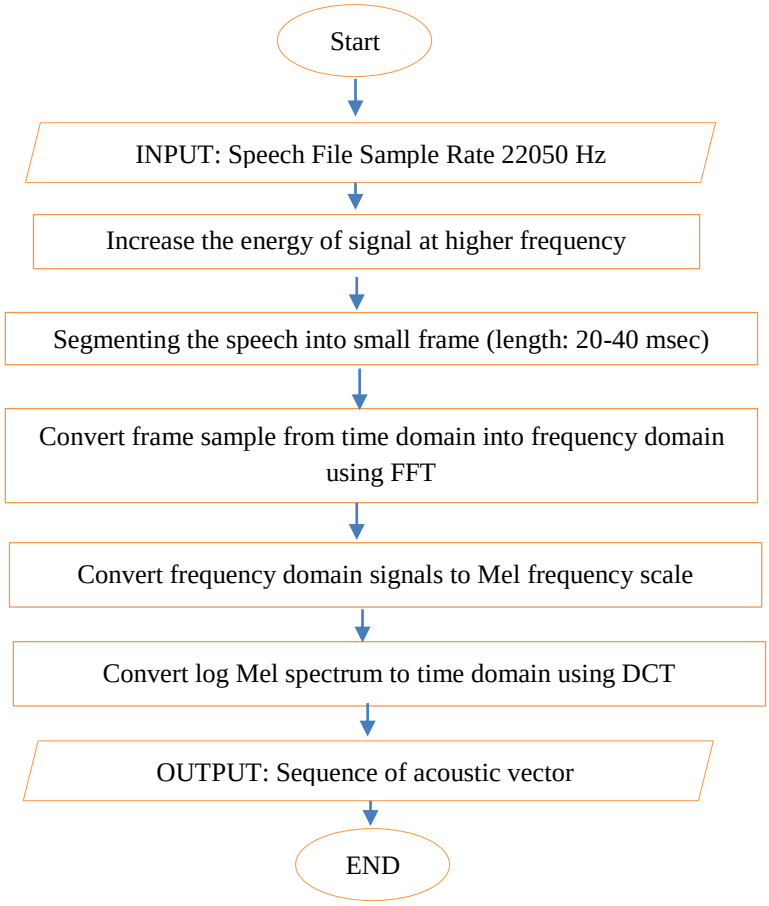


Figure 3.4. MFCC feature extraction technique flowchart (Singh et al., 2012).

The MFCC steps were implemented using MATLAB in order to employ feature extraction phase. Figure 3.5 shows the used MATLAB code in feature extraction phase.

```
%% FEATURE EXTRACTION
    % Preemphasis filtering
   speech = filter( [1 -alpha], 1, speech );
   % Framing and windowing (frames as columns)
   frames = vec2frames( speech, Nw, Ns, 'cols', window, false );

   % Magnitude spectrum computation (as column vectors)
   MAG = abs( fft(frames,nfft,1) );
   % Triangular filterbank with uniformly spaced filters on mel scale
   H = trifbank( M, K, R, fs, hz2mel, mel2hz ); % size of H is M x K

   % Filterbank application to unique part of the magnitude spectrum
   FBE = H * MAG(1:K,:); % FBE( FBE<1.0 ) = 1.0; % apply mel floor

   % DCT matrix computation
   DCT = dctm( N, M );

   % Conversion of logFBEs to cepstral coefficients through DCT
   CC =  DCT * log( FBE );

   % Cepstral filter computation
   lifter = ceplifter( N, L );

   % Cepstral filtering gives filtered cepstral coefficients
   CC = diag( lifter ) * CC; % ~ MFCCs
```

Figure 3.5. the MATLAB code of feature extraction phase using MFCC

Basically, in the first step we provide an audio file for a test sentence which was recorded before in order to implement framming stage. For instance, we provide a speech audio file for a sentence containing eight words ('*To men whom want to quit work someday*') which was recorded by one person and the system was trained by another person. Hence, the system was capable to provide eight speech signals for each sentence word. Figure 3.6 shows the signal of the ('*Whom*') for testing sides. Speech signals for each word were privewed in Appendix I.

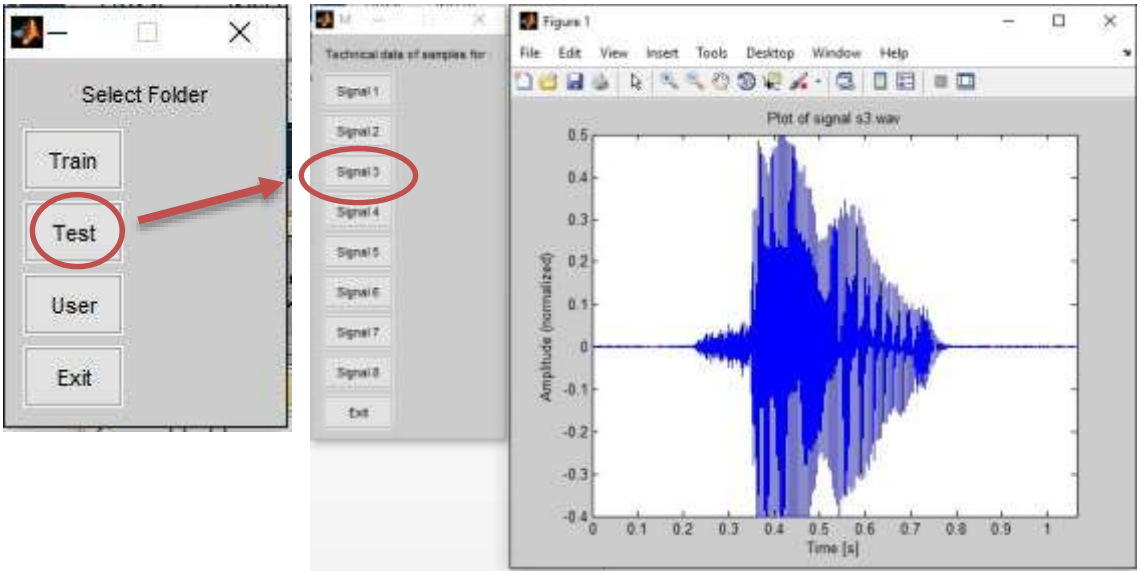


Figure 3.6. Framing process for feature extraction stage using MATLAB – an example of speech signal for the word ('whom') from the test sentence – in the test side.

Figure 3.7 shows the training side for the word ('whom') in the training side in the test sentence ('*to men whom want to quit work someday*').

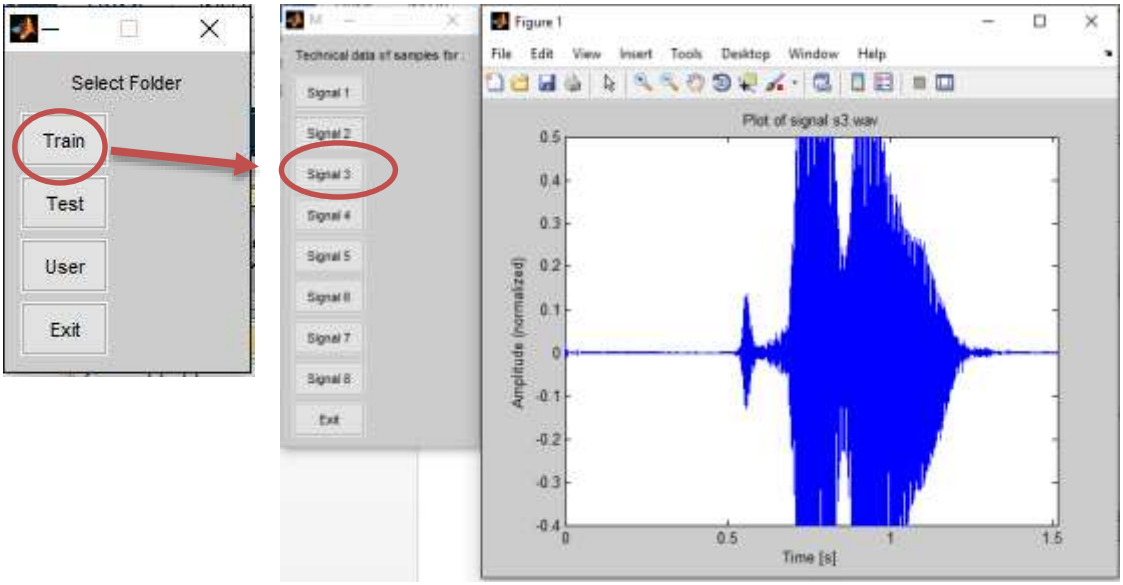


Figure 3.7. Framing process for feature extraction stage using MATLAB – an example of speech signal for the word ('whom') from the test sentence – in the training side.

Figure 3.8 shows the process of generating the power spectrum (i.e. magnitude spectrum) for the test sentence speech audio file that was consisted from 100 sample frame.

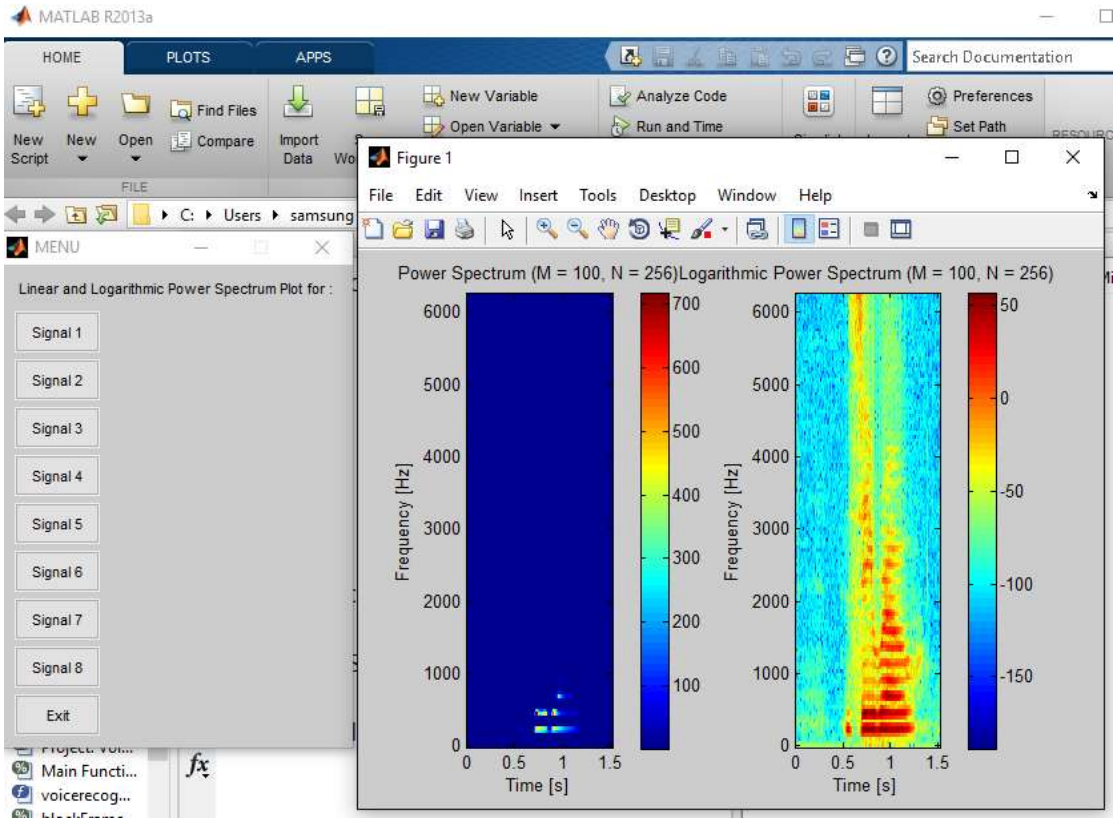


Figure 3.8. The power spectrum for the third word in the speech signal that was segmented into 100 sample frame

Figure 3.9 shows the power spectrum after increasing the number of frames for the third speech signal in the test sentence. The frames number was increased as 108, 220,438 frame respectively to prove doubling and troubling the number of frames.

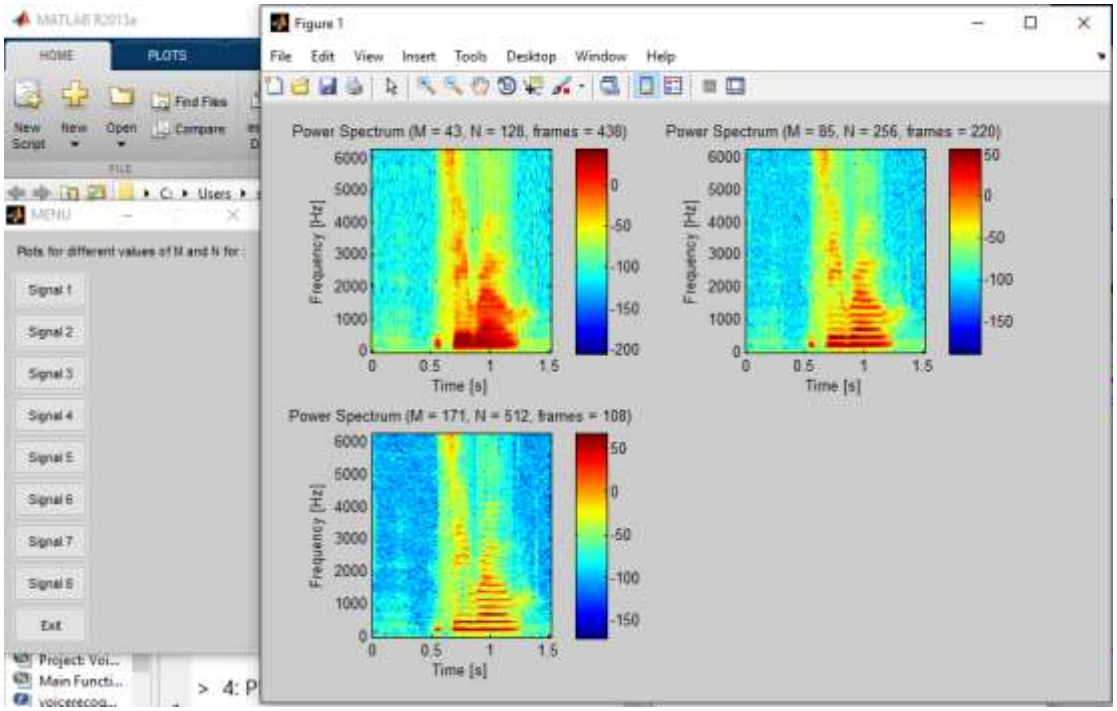


Figure 3.9. The power spectrum after increasing the number of frames for the third signal in the test sentence ('*to men whom want to quit work someday*') speech audio file.

The Mel frequency filter bank is generated based on the theoretical background that was mentioned in chapter two. Figure 3.10 shows the Mel-frequency filter bank of the speech audio file for the test sentence.

The result of MFCC is the acoustic model which represents the feature extraction of the speech audio file. Figure 3.11 shows the acoustic model that was generated after testing the audio speech file for the second word ('men') and the third word ('*whom*')' in the test sentence ('*to men whom want to quit work someday*').

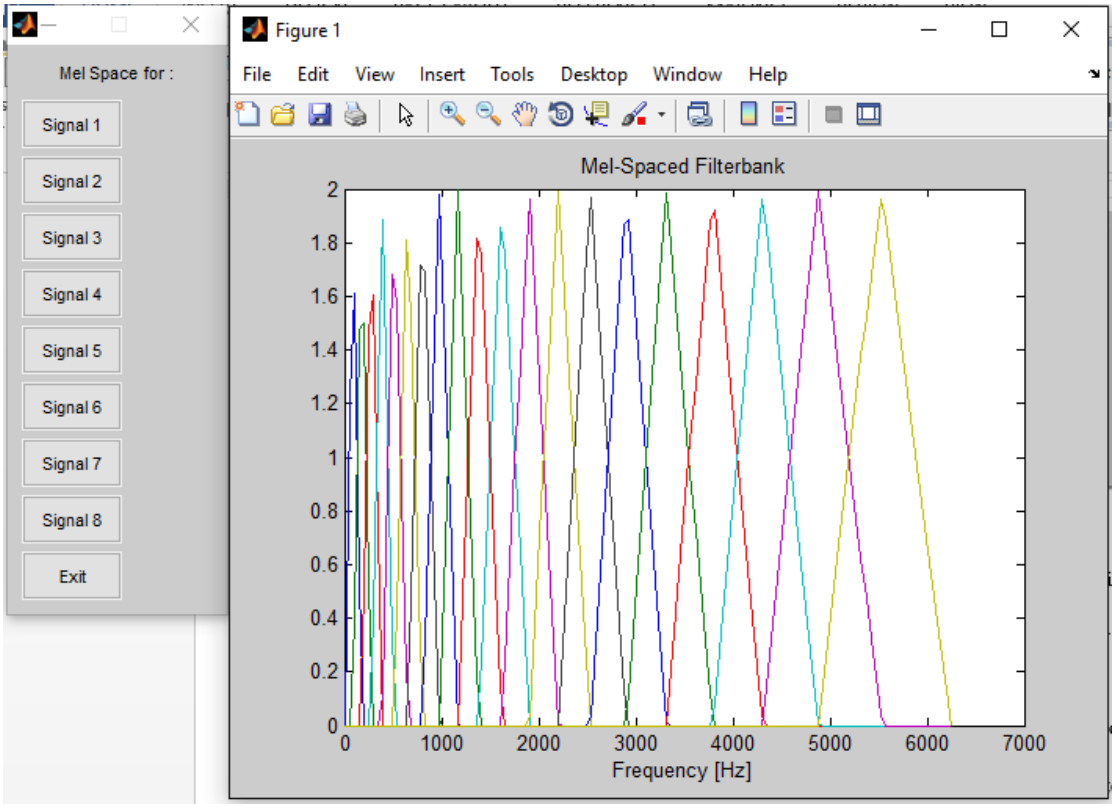


Figure 3.10. The Mel-Frequency filter bank of the third word ('*whom*') in the test sentence

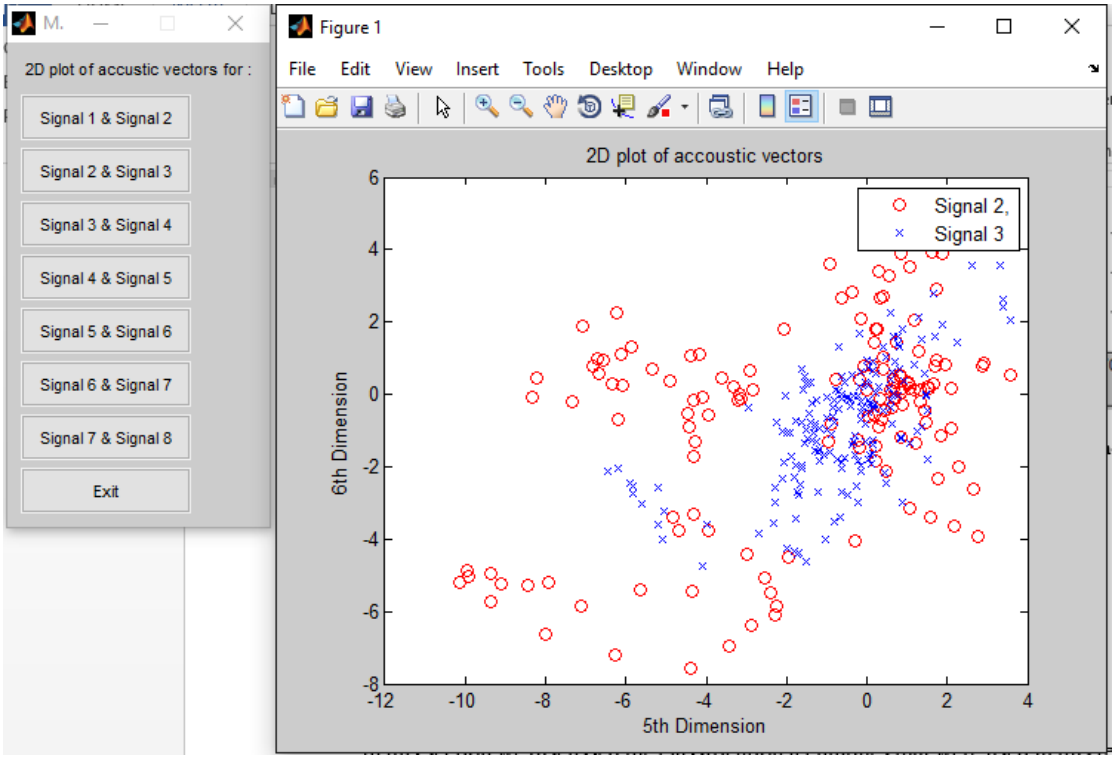


Figure 3.11. The acoustic feature plot of the second word ('men') and the third word ('whom') in the test sentence

The acoustic vector that was generated between the third word ('*whom*') and the fourth word ('*want*') is shown in Figure 3.12

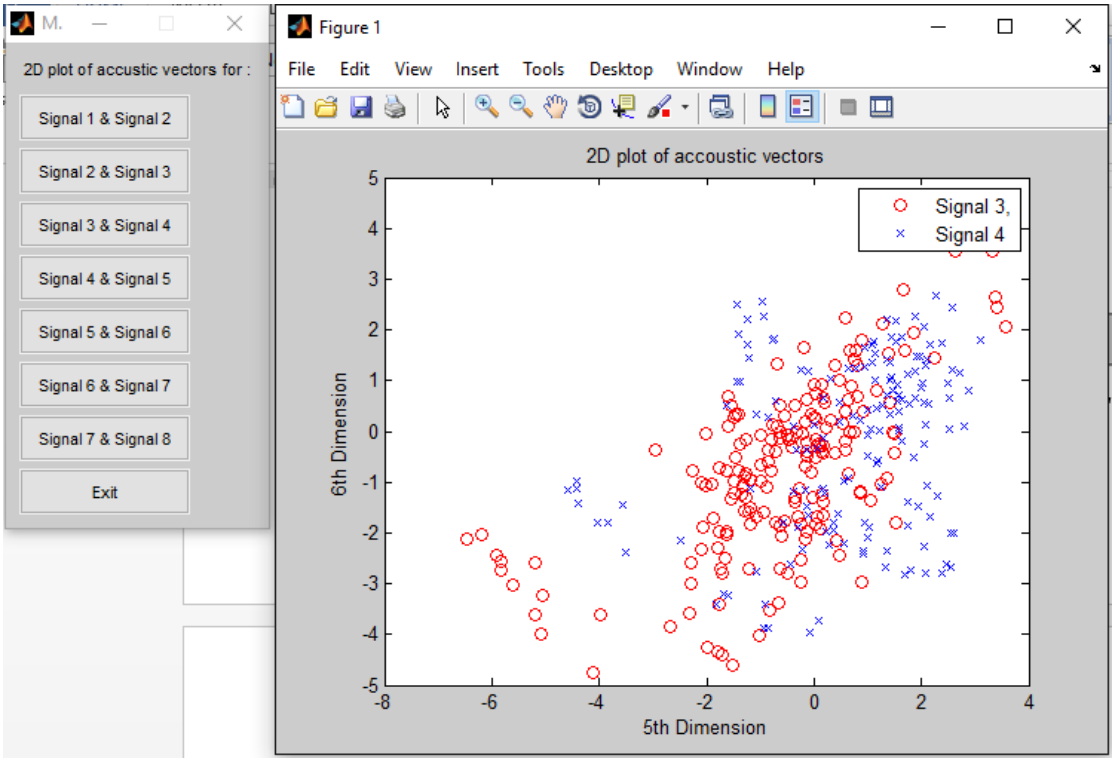


Figure 3.12. The acoustic vector between the third word signal ('whom') and the fourth word ('quit')

The coordinates of the acoustic vector that was generated is shown in Figure 3.13 that shows a part of the table in MATLAB.

MFCCs × time × time_frames × FBEs ×

FBEs <20x265 double>

	1	2	3	4	5	6	7	8	9	10
1	67.5528	108.3720	113.9487	149.2825	222.7137	223.1379	66.6868	51.7321	39.6631	84.0876
2	167.6184	90.3636	144.6622	126.9550	150.4947	124.5623	64.1506	51.1115	49.8183	123.5120
3	105.3096	58.7759	137.1790	122.5051	182.7674	96.8566	91.8998	57.4100	57.2935	95.0212
4	98.4059	79.2347	32.3319	120.6143	167.1641	63.2897	77.4032	91.2468	72.8558	43.8582
5	107.6248	65.4317	72.7092	90.2046	120.7485	64.9553	88.2160	112.9709	96.9116	57.4375
6	87.5820	137.9309	149.9988	82.6124	125.8934	105.3958	119.6077	141.3874	73.1361	47.6269
7	292.4529	280.0751	273.3245	169.6460	179.8264	130.2333	149.1866	107.3998	104.7796	105.1650
8	226.9329	200.7275	194.8294	100.7338	137.3134	212.3182	133.5340	133.0243	133.1818	119.0717
9	237.3539	206.6946	159.8696	169.7976	199.9292	238.9078	170.5831	164.5643	155.7250	135.4924
10	481.2276	655.4453	829.8495	520.9458	445.4729	390.4712	279.9079	266.7439	252.9535	187.1827
11	404.6871	881.0064	726.2837	373.9614	372.7432	426.2475	393.6457	211.4397	256.5289	219.0084
12	766.3988	769.4361	851.6162	569.5231	644.9193	743.7728	461.2920	496.2487	304.5218	197.5632
13	634.1303	820.0026	952.3491	779.6609	675.1041	397.8774	533.1329	381.3633	252.3530	208.5933
14	197.1049	435.1489	245.7852	332.0085	301.6386	209.5055	169.3977	143.2201	190.3862	139.2028
15	237.2120	364.3112	291.5853	339.5644	205.5712	276.2341	290.1047	129.5664	144.9901	224.2096

Figure 3.13. Part of the MFCC acoustic vector coordinates between two speech signals from the conducted experiment

3.2.3. Feature Extraction classification techniques

In this section we discuss the classification techniques that were used in this research. Hence, this research took into consideration three common classification techniques which were HMM, DTW, and DBN.

The implementation of the DTW in this experiment was employed based on the discussion in section theoretical background and related work in chapter two. Thus, the experiment was conducted by taking two MFCC acoustic vectors which were the acoustic vectors for both the training and testing of speech audio file for the third word ('whom') and the fourth word ('want'). Figure 3.14 shows the MATLAB code of DTW function used in speech recognition process in our experiment.

```
function d=dtw(s,t,w)
% s: signal 1, size is ns*k, row for time, column for channel
% t: signal 2, size is nt*k, row for time, column for channel
% w: window parameter
%      if s(i) is matched with t(j) then |i-j|<=w
% d: resulting distance
if nargin<3
   w=Inf;
end
ns=size(s,1);
nt=size(t,1);
if size(s,2)~=size(t,2)
   error('Error in dtw(): the dimensions of the two input signals do not match.');
end
w=max(w, abs(ns-nt)); % adapt window size
%% initialization
D=zeros(ns+1,nt+1)+Inf; % cache matrix
D(1,1)=0;
%% begin dynamic programming
for i=1:ns
   for j=max(i-w,1):min(i+w,nt)
      oost=norm(s(i,:)-t(j,:));
      D(i+1,j+1)=oost+min( [D(i,j+1), D(i+1,j), D(i,j)] )
   end
end
d=D(ns+1,nt+1);
```

Figure3.14. the MATLAB code using for implementing Data Time Wrapping (DTW) for two feature vectors using MFCC

Consequently, we can see from figure 3.15 the similarity between two vectors of training and testing acoustic vector based on the dark stripe which represent a high similarity values down the leading diagonal. Figure 3.15 shows the time wrapping between the test and training acoustic vectors.

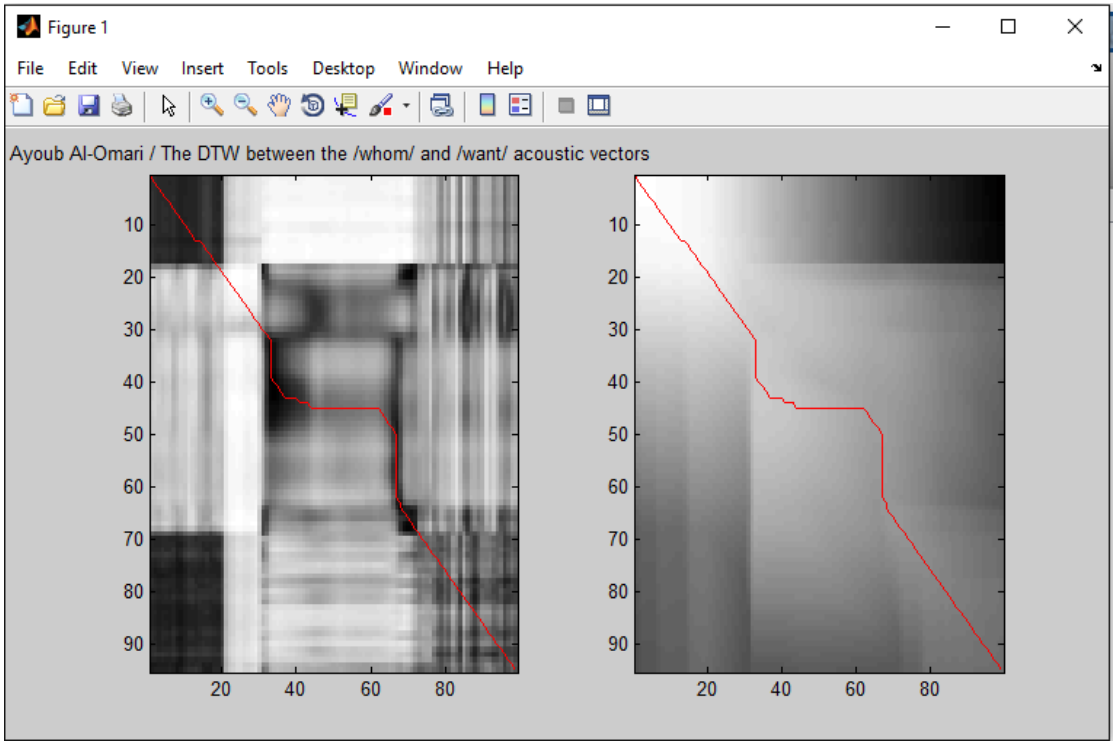


Figure 3.15. The similarity values after applying DTW on two acoustic vector in the proposed experiment

However, we applied the second classification technique (i.e. HMM) in order to study the results of the speech recognition based on the acoustic features vector from MFCC phase. Figure 3.16 shows the training phase of HMM system in this research using MATLAB environment.

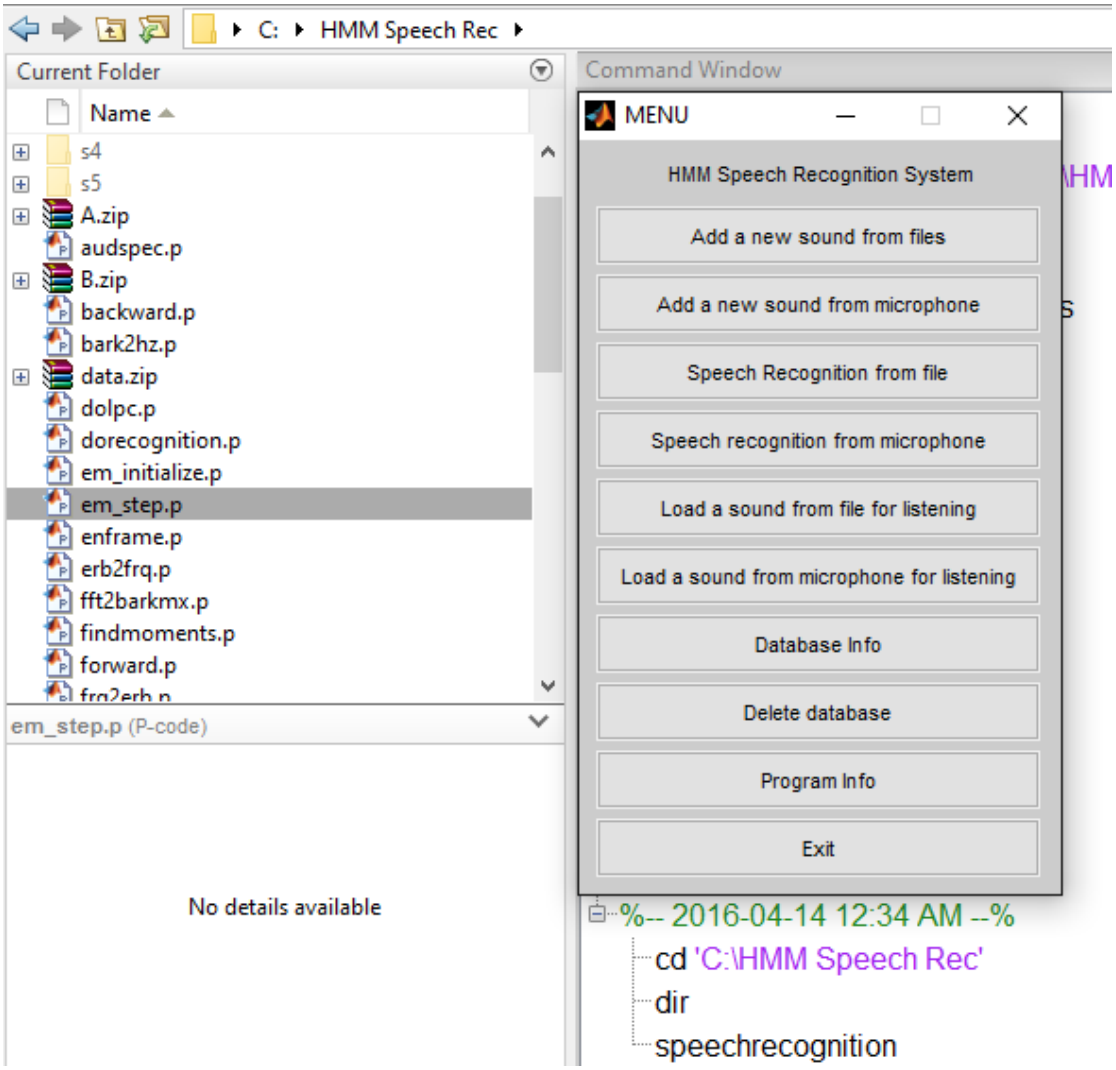


Figure 3.16. the main GUI of HMM application for training and recognition algorithm for a speech as well as for data collection phase in this research

The training algorithm in this research concentrated on recording speaker's voice while talking a specific sentences (Appendix A). These sentences were used in each environmental combination (i.e. feature extraction and classification technique. Figure 3.17 shows the training process for HMM system.

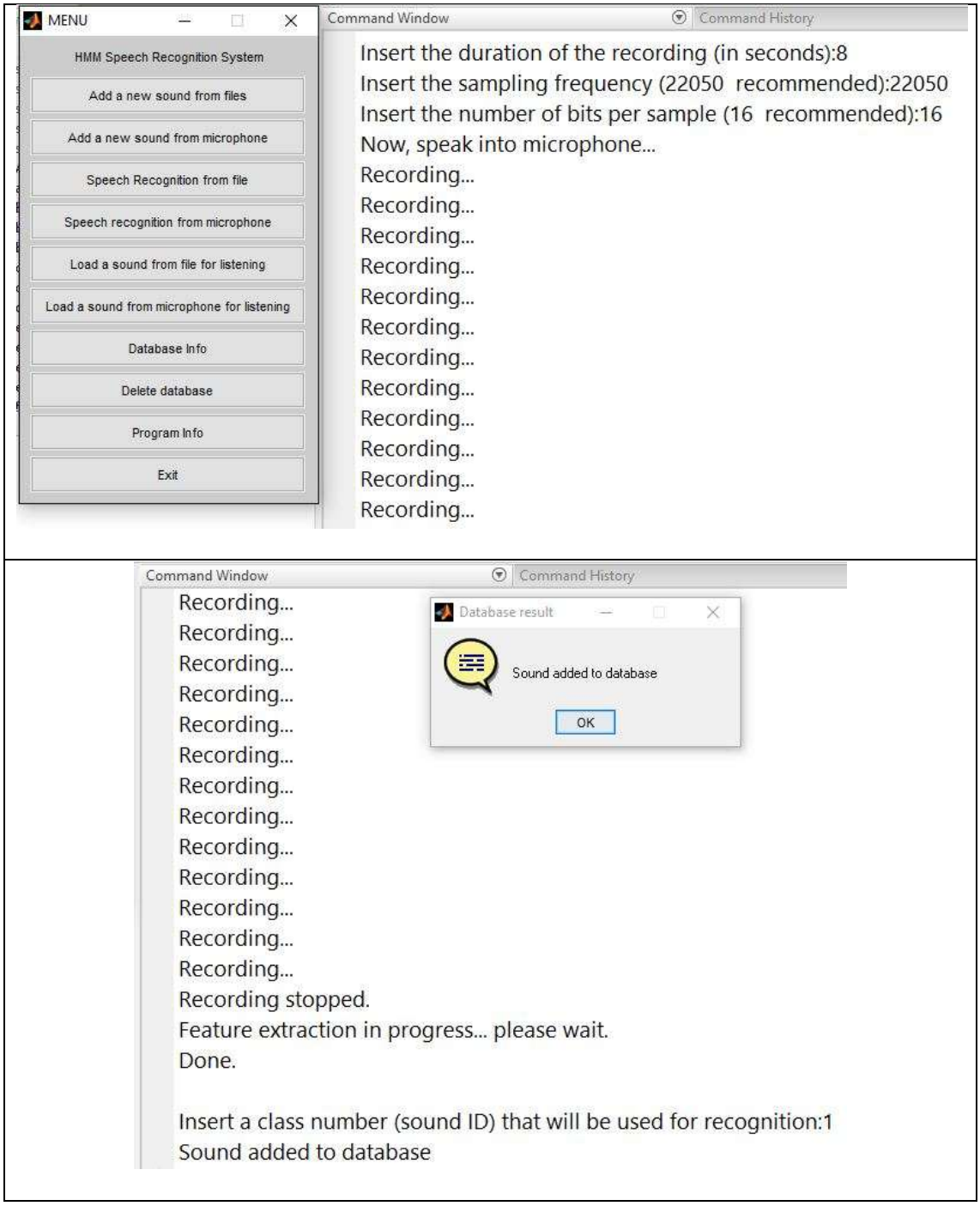


Figure 3.17. training and recognition algorithm in HMM using MATLAB

In order to evaluate the speech recognition process we add a number for the sound file from the training set. Thus, we made the record process to implement a test side and the system is responsible to retrieve the ID number of the speech audio file from the

training set that is similar to the target speech. Figure 3.18 shows the results of HMM speech recognition process.

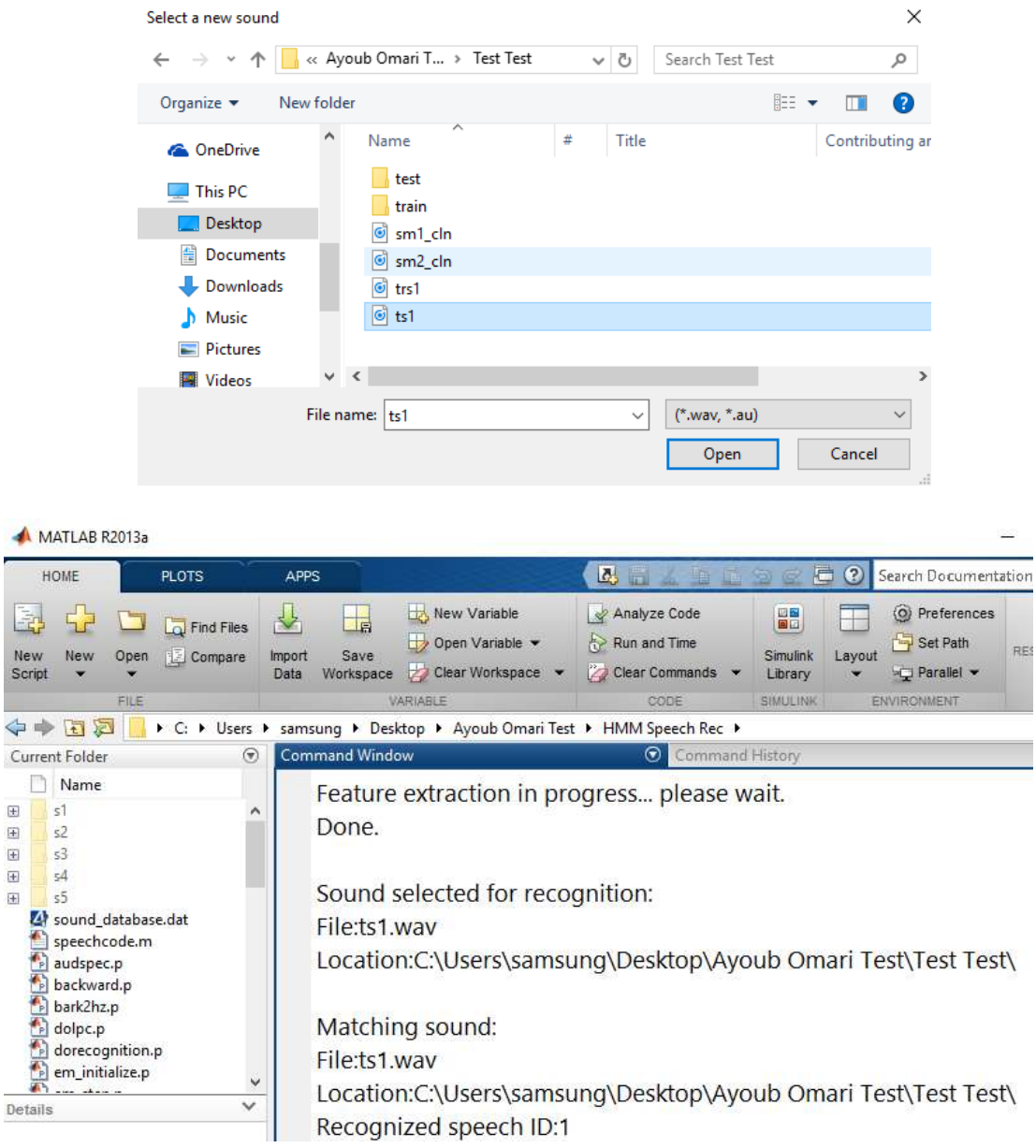


Figure 3.18. The matching process in speech recognition after selecting the recorded file to be compared with the trained file ts1.wave

.

3.2.4. Machine Translation (MT) phase

Several MT engines are available on the internet to provide a translation services from any input language into a target one. For instance, Google Translate, Bing Translate, Yahoo Babelfish, and Systran. Thus, these MT engines uses different types of translation models such as rule-based translation, statistical-based translation, and hybrid-based translation models.

Basically, in this research we investigated the general framework of each translation model in order to make a comparison between them. Hence, we applied the results from the speech recognition phase to feed the MT phase in order to capture the effects of classification techniques in ASR on the translation model in term of accuracy of the translated text from English to Arabic language.

A. Rule-based Translation Model

The core idea of this translation model is to build a production rules and a dictionary for each language pairs. Then, the MT parses the text that has to be translated to show the results in the target language. Thus, any rule based translation system contains three main processes which are syntactic process, semantic process, and morphological analysis that uses a large set of rules. Figure 3.19 illustrates the flow chart of rule based translation systems for applying rules process to identify the position of the words.

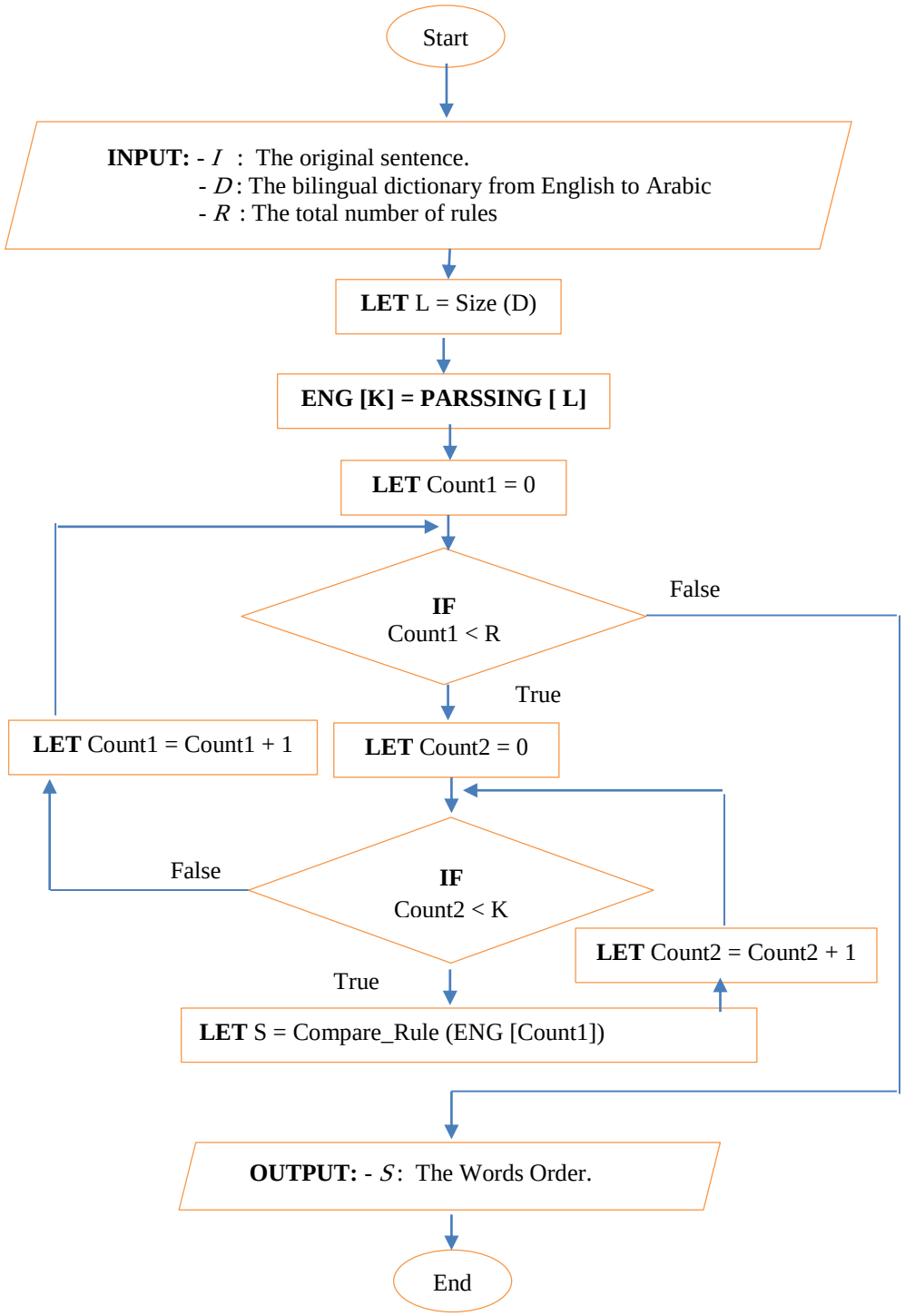


Figure 3.19. Flowchart of rule based translation model for the applying rules process to find word's position in the sentence before finding the word's meaning (Rhaman and Tarannum, 2012).

Furthermore, the rule-based translation model compares each word from the original text with the words in the language dictionary to find its meaning. Figure 3. 20 illustrates the flowchart of rule based translation model of finding words meanings process.

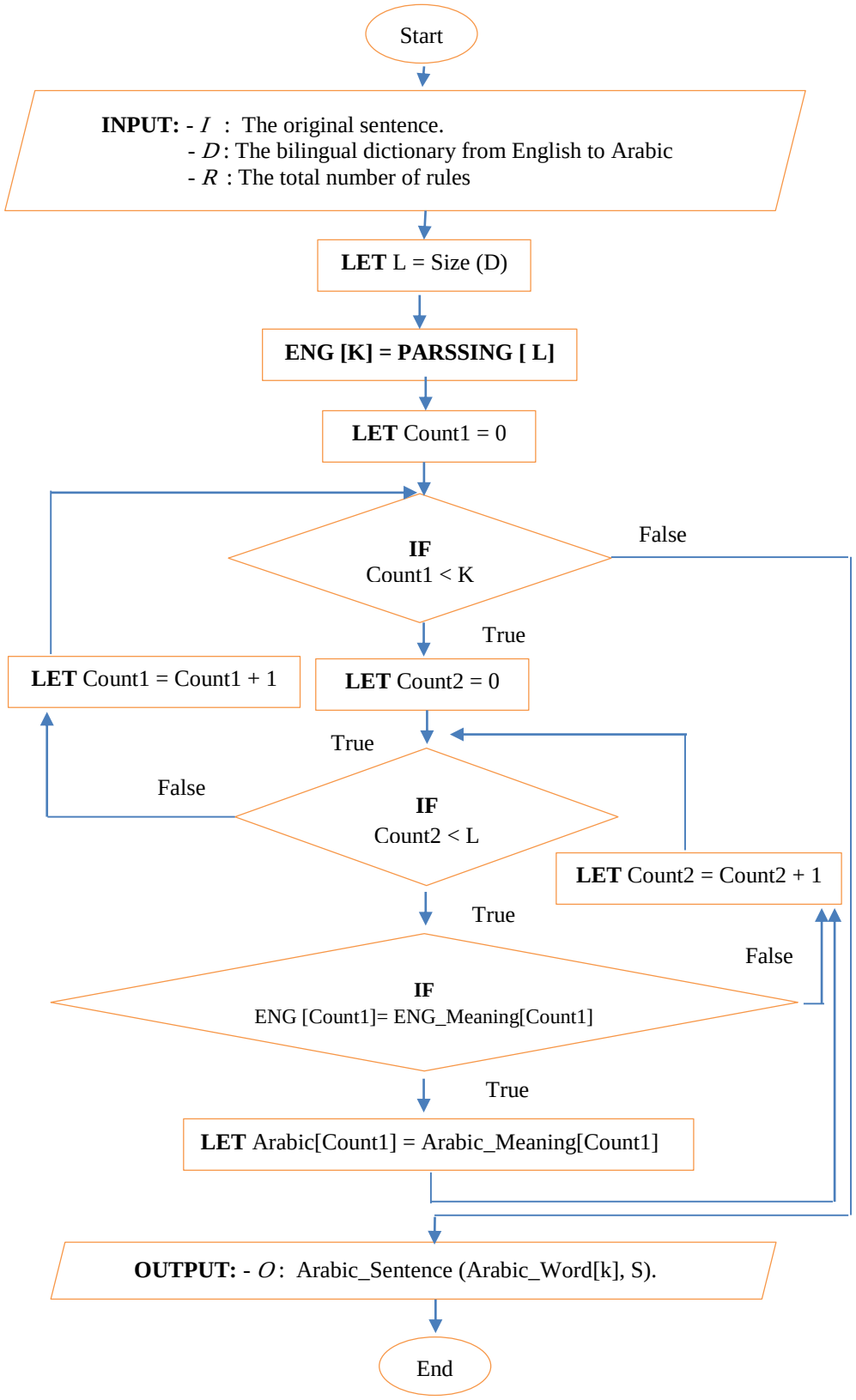


Figure 3.20. Flowchart of rule based translation model in finding word's meaning to translate the input text from English to Arabic (Rhaman and Tarannum, 2012).

B. Statistical-based Translation Model

Basically, the conditional probability is used in the statistical based translation model to describe the correspondence between two sentences (i.e. the input and target sentences). Thus, for a given source language sentence $F_1^J = f_1, f_2, f_3 \ldots, f_j$ that should be translated into a target language sentence $e_1^J = e_1, e_2, e_3 \ldots, e_I$ and by benefit from the log linear approach that uses the maximum entropy framework to search for the best translation of the input sentence, the translation decision rule is in (equation 12) (Och and Ney, 2002).

$$e^{\wedge I}{}_1 = \arg\max \sum_{m=1}^{M} \lambda_m h_m \left(e^I{}_1, f^{\ J}{}_1 \right) \qquad (12)$$

Where:

- λ_m : is the weight of feature function
- $h_m \left(e^I{}_1, f^{\ J}{}_1 \right)$: The translation and language model probabilities.

The statistical based translation model is consisted into two sub models which are the (1) language model that is used to describe the correctness of the target language sentence, and (2) the translation model which contains two sub models which are the lexicon and alignment models that is responsible to translate the source language sentence *F* into a target language sentence *e* by maximizing a linear combination and weights between them. Figure 3.21 illustrates the block diagram of statistical based translation system and its main components.

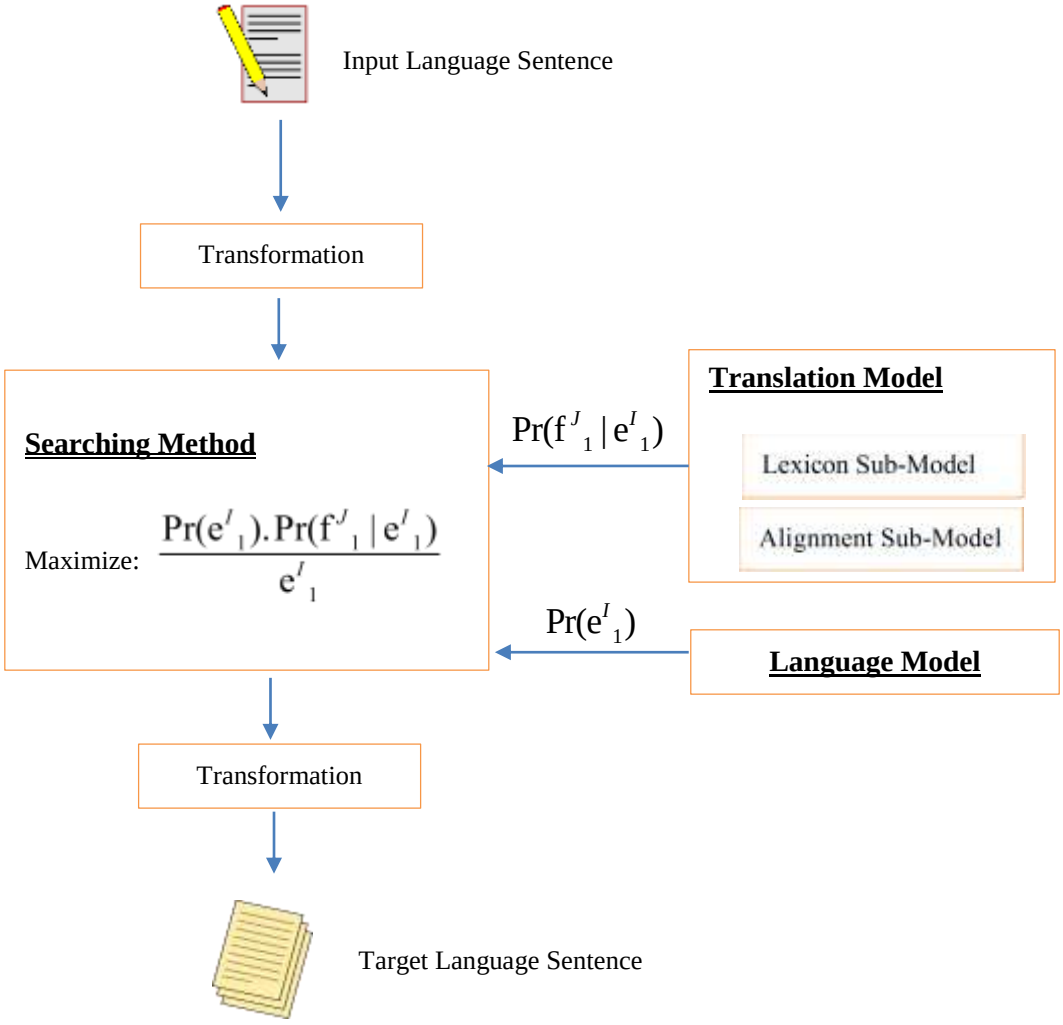


Figure 3.21. The block diagram of statistical based translation model showing its main processes and sub-models (Kazuma et al., 2011)

In the alignment sub model a Viterbi algorithm (i.e. a dynamic programming algorithm for finding the most likely sequence of hidden states) is employed to show the index a^J_1 of the word in e^I_1. Hence, the index could be computed based on (equation 13) (Och and Ney, 2002).

$$a^J_1 = \arg\max Pr(f^J_1 a^J_1 | e^I_1) \quad (13)$$

Consequently, in order to make a comparison we had examined three available online MT engines that are available for free on the internet to translate from English to

Arabic language. Thus, the selection decision was based on the MT model used in each engine. Therefore, we had chosen Google Translate (http://translate.google.com) which uses statistical based translation model, IBM API language Translate (http://language-translation-demo.mybluemix.net/) which uses rule based translation model, and Systran Translation (http://www.systransoft.com/) which uses hybrid based translation model that combines statistical and rule based translation models.

3.3. Evaluation metrics

In this context, the Word Error Rate (WER) is calculated by finding the number of inserted words in the target sentence *I*, the number of the deleted words in the target sentence *D*, the number of the substituted words in the target sentence *S*, and the number of the correct words in the target sentence *C*. Thus, the WER measure is calculated based on (equation 14) (Park et al., 2008). Figure 3.22 shows an example of finding WER for a speech news file.

$$WER = 100 \times \frac{(S + D + I)}{N} \tag{14}$$

Where:

N = D + S + C

Furthermore, the Word Recognition Rate (WRR) metric was represented as the complement of the WER. Therefore, we calculated WRR based on the equation (15) (Park et al., 2008).

$$WRR = 1 - WER \tag{15}$$

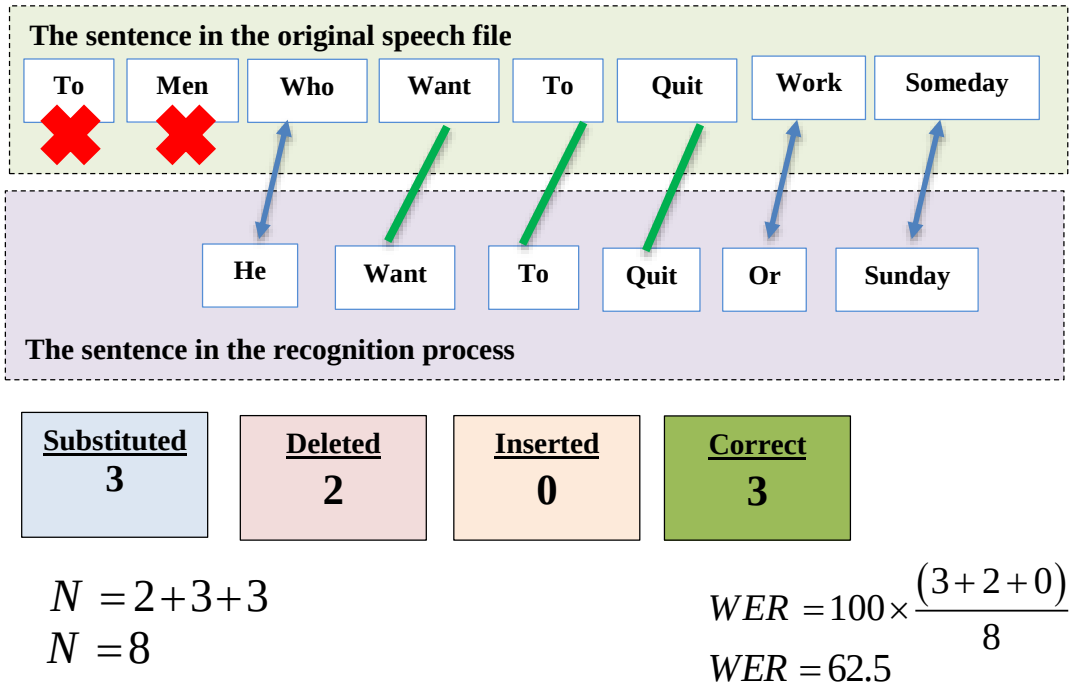


Figure 3.22. Example of calculating the accuracy measure WER for a sentence from the news category.

CHAPTER FOUR

Experimental Results

4.1. Overview

This chapter discusses the results of the proposed experiments which were discussed in chapter three. In this chapter, the performance and accuracy of ASR and MT engine were discussed. Furthermore, the results discussion in this chapter covered the proposed levels in this study. Therefore, the flow of results in this chapter was discussed as the following:

- The experimental results of the Word Recognition Rate (WRR) in ASR level (i.e. classification technique phase) were discussed.
- The experimental results of the Word Error Rate (WER) in MT level (i.e. matching models) were discussed.

4.2. The experimental results of ASR level

In this section, we provided the experimental results of comparing the speech recognition results after employing three classification techniques which are the DTW, HMM, and DBN techniques for different types of speech. Therefore, the experiments in this research took into consideration four types of speech which are news, scientific phrases, conversational, and control phrases. Table 4.1 shows the sample speech sentences of news speech category. Thus, we examined each speech category with the three matching models by computing WRR to identify the accuracy of speech recognition.

Table 4.1. The news speech file sample sentences

Number	The Test Sentence	Category
1	Things You Can Learn From the Apple Store	News
2	Do You Do Any of These Ten Embarrassing Things?	News
3	To Men Who Want to Quit Work Someday	News
4	Why Facebook is Making Life Hard	News
5	Where You Can Go in a Good Used car	News
6	Welcome to the New Civil War	News
7	The Child Who Won the Hearts of All	News
8	Suicide of a Hacker	News
9	How to Publish a Book	News
10	Are You Ever Tongue-tied at a Party?	News
11	Discover the Fortune that Lies Hidden in Your Salary	News
12	Do You Make These Mistakes in English?	News
13	How I Improved My Memory in One Evening	News
14	How to Win Friends and Influence People	News

Consequently, we applied the ASR system with DTW approach as a speech matching approach. Table 4.2 shows the results of speech recognition phase for news category. Therefore, we calculated the accuracy of the ASR system by finding the WER for each experiment.

Table 4.2. The recognition results of applying news speech files in ASR system uses DTW

Number	Recognition Results (MFCC + DTW) ASR	Category
1.	Things you can learn from their history.	News
2.	Do you do anything used embarrassing things.	News
3.	He didn't want to quit or Sunday.	News
4.	It is just breaking my heart.	News
5.	Where you can go in and who's gonna.	News
6.	Welcome to the new server or.	News
7.	Did Charles who won the hearts of four.	News
8.	Suicide of Hanukkah.	News
9.	How to publish a book.	News
10.	Are you trying to tie the two party.	News
11.	Discover the first tune that lies hidden in your salad.	News
12.	Do you make these mistakes in English.	News
13.	However improved my memory in one evening.	News
14.	How to win friends and influence people	News

Table 4.3 shows the accuracy results in term of WER measure after applying news category speech files in an ASR system that employs the DTW matching approaches.

Table 4.3. The WER parameter results after analyzing the news speech category using an ASR system uses DTW.

Sentence No.	WER
1.	42.857143
2.	42.857143
3.	71.428571
4.	42.857143
5.	50
6.	33.333333
7.	50
8.	33.333333
9.	0
10.	71.428571
11.	37.5
12.	0
13.	22.222222
14.	0
Average WER for ASR using DTW	**35.55839**

Consequently, the DTW matching approach achieved 35.5% as an average WER via applying news speech files. Hence, the second stage in this level is to find the WER for ASR system that uses HMM matching approach. Table 4.4 shows the WER results after analyzing the news speech category using an ASR system that employs HMM for matching criteria.

Table 4.4. The WER parameter results after analyzing the news speech category using an ASR system uses HMM.

Sentence No.	WER
1.	0
2.	71.428571
3.	33.333333
4.	33.333333
5.	0
6.	33.333333
7.	0
8.	87.5
9.	0
10.	88.888889
11.	87.5
12.	0
13.	12.5
14.	0
Average WER for ASR using HMM	**31.9569**

In the third stage in this level, we examined the results in term of WER on an ASR system that employs DBN as a matching approach. Table 4.5 shows the results of WER in an ASR system that uses DBN matching approach. The detailed results of the first, second, and third stages are found in Appendices A, B, and C respectively.

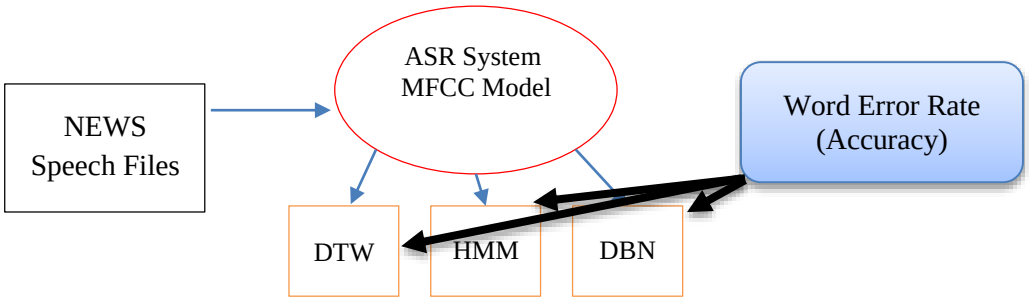


Figure 4.1. The main processes diagram of the news speech files and comparison experiments

Table 4.5. The WER parameter results after analyzing the news speech category using an ASR system uses DBN.

Sentence No.	WER
1	0
2	11.11111
3	12.5
4	16.66667
5	25
6	80
7	12.5
8	33.33333
9	0
10	42.85714
11	36.36364
12	14.28571
13	12.5
14	0
Average WER for ASR using DBN	**21.222**

4.2.1. ASR phase results discussion

In this section we provide a discussion for the results in ASR system. Thus, the results showed that the using DTW matching approach achieved the highest error rate in new category which represented with the percentage (35.9%). Furthermore, the ASR system that used HMM matching approach showed a little difference in error rate compared with DTW for the same speech files category. In contrast, the ASR system with DBN matching approach achieved the best result that made a drastic gap in error rate since it is the lowest result. Figure 4.2 shows a graph that represent the averages of WER in ASR system using the same feature extraction model (i.e. MFCC) within three different matching approaches DTW, HMM, and DBN respectively.

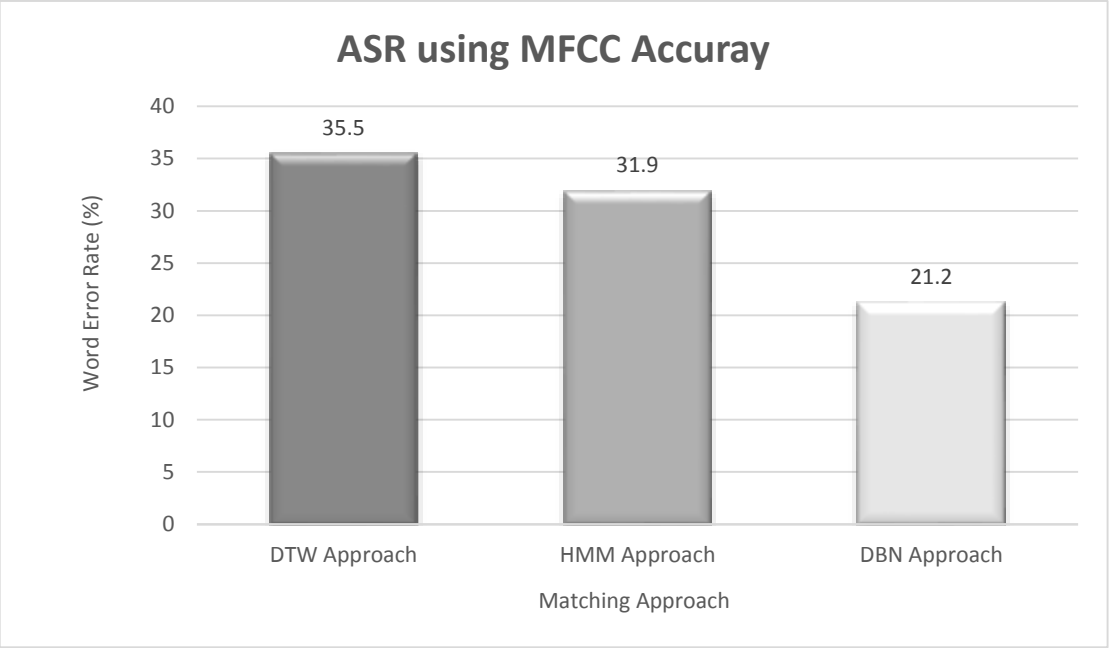


Figure 4.2. The averages of WER chart of ASR systems using DTW, HMM, and DBN approaches.

A group of test for different audio speech intervals was performed to estimate the accuracy of each environment by computing the WER of each ASR environment. Thus, in order to ease the comparison the recorded speech files were classified into four categories news category (i.e. which was mentioned in the previous section), the conversational category, the scientific phrases category, and controls category. These categories were chosen to cover the most used application for different fields that need speech recognition techniques.

Consequently, the same ASR environmental components choices were used to made the experiments after applying the recorded conversational category audio files. Table 4.6 shows the results of accuracy for each audio file in the conversational category in term of WER.

Table 4.6. results of accuracy for each audio file in the conversational category in term of WER.

Conversational			
Audio File #	**DTW**	**HMM**	**DBN**
15	0	0	63.4
16	0	0	0
17	0	0	23.5
18	0	2	3.9
19	66.66667	33.3	15.3
20	88.88889	68.1	54.8
21	0	0	0
22	0	0	13.9
23	28.57143	22.2	0
24	55.55556	16.7	0
25	28.57143	11.3	21.9
26	0	8.9	22.5
27	43.8	42.11	36.17
28	0	0	0
29	50	34.15	51.3
Averages	**24.13693**	**15.91733**	**20.44467**

Based on the results shown in Table 4.6 and in order to ease the comparison between ASR matching techniques we provided a graphical chart of the accuracy averages in term of WER for each matching technique. Figure 4.3 shows a chart of accuracy averages for conversational audio files.

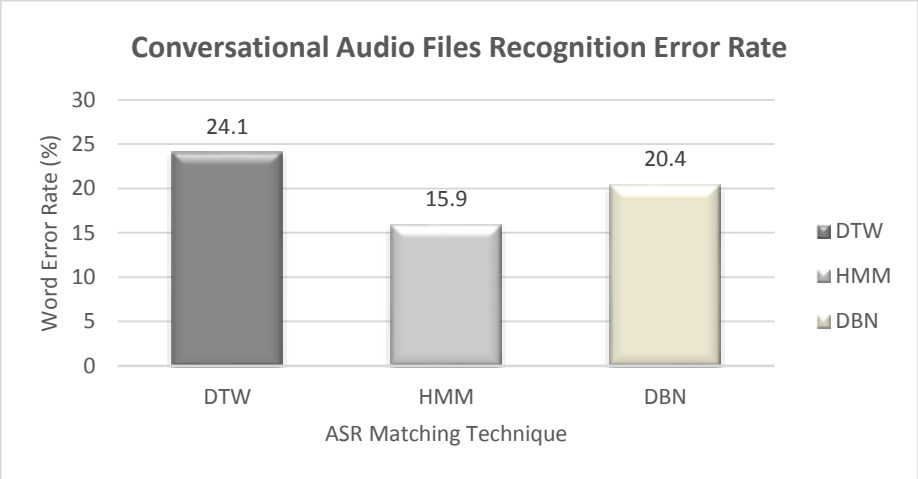


Figure 4.3. shows a chart of accuracy averages for conversational audio files.

From conversational category audio files point of view, the results showed that each matching technique was achieved different rate value noticed that these audio files were recorded by the same person and in the same environment (i.e. the same devices and computer's software). Actually, the ASR system that used the HMM matching technique achieved the best accuracy compared with other ASR systems. Thus, the ASR system with HMM achieved the lowest WER which was 15.917%. However, the ASR with both DTW and DBN systems were achieved 24.13% and 20.44 % respectively, which made a drastic gap in error rate compared with ASR with DBN matching technique.

Therefore, we found the WRR (i.e. the complement of WER) which was used to determine the recognition rate of the system's accuracy by converting speech into text. Thus, WRR showed that in case of using conversational category audio file the most accurate combination was by using MFCC with HMM matching technique. On the other hand, for the same audio files category the worst accuracy results was achieved by using the combination of MFCC and DTW. The results of WRR for ASR systems with DTW, HMM, and DBN were 75.87%, 84.1%, 79.6% respectively. Figure 4.4 shows the accuracy chart in term of WRR for conversational category audio files.

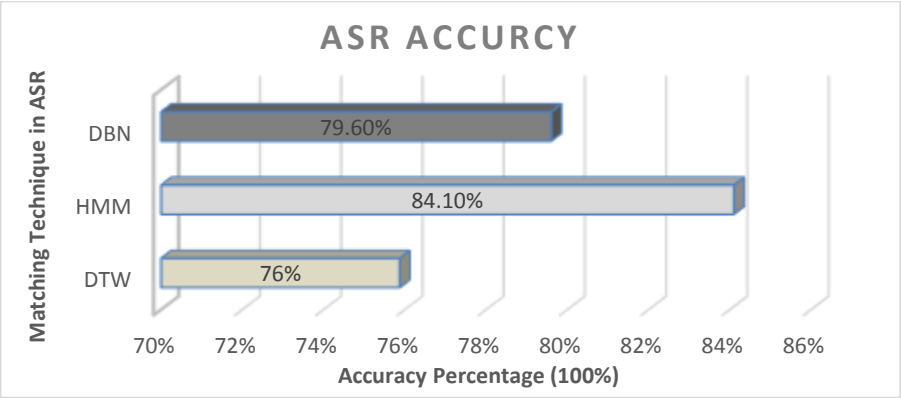


Figure 4.4. Accuracy percentage chart for each system of ASR based on WRR

In this context, we also extracted the results of applying the same matching techniques on the scientific phrases category in order to extract the WER. Table 4.7 shows the WER results were calculated after testing the scientific phrases audio files.

Table 4.7. The WER results of testing scientific phrases category audio files in ASR systems with DTW, HMM, and DBN.

Scientific Phrases			
Audio File #	**DTW**	**HMM**	**DBN**
30	0	0	2.7
31	42.5	0	29.17
32	8.5	8.5	44.7
33	12.9	14.8	63.1
34	62.07	62.07	56.6
35	9.3	11.63	0
36	2.3	4	0
37	25.5	25.5	40
38	0	2.86	8.5
39	2.8	10.5	48.2
Averages	**16.587**	**13.986**	**29.297**

For more clarification, a graphical chart was drawn to view the averages of accuracy parameter for the three matching techniques. Figure 4.5 the WER averages of testing the scientific phrases using the three ASR environments.

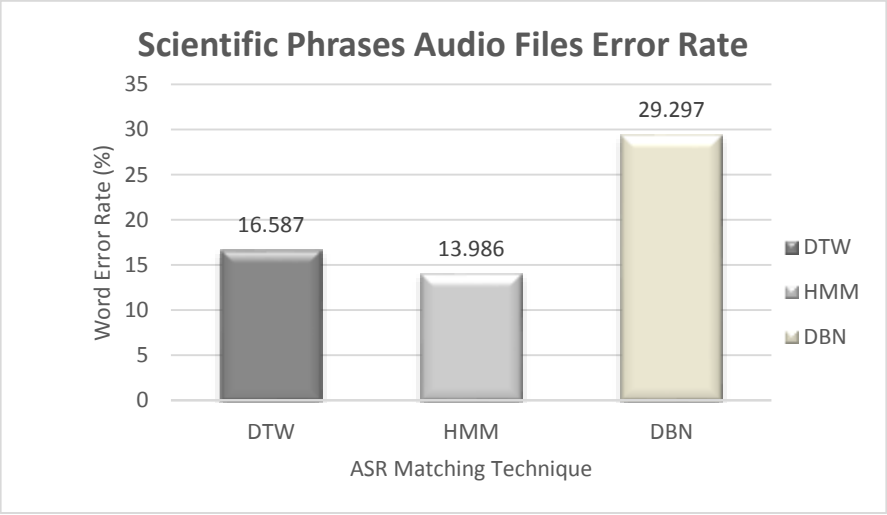


Figure 4.5. the WER averages of testing the scientific phrases using the three ASR environments.

As shown in Figure 4.6, the results showed an occurrence of diversity between the three ASR systems in term of error rate for the scientific speech audio files. Hence, the HMM achieved for the second time the best WER result with 13.9% compared with the DTW, and DBN. Furthermore, a drastic gap existed via DBN matching technique which achieved the worst results with 29.3% in term of WER too. Therefore, the accuracy of the ASR system in case of testing scientific audio files was computed based on WRR. Thus, the accuracy results were found for DTW, HMM, and DBN as 83.4%, 86%, and 70.7% respectively. Figure 4.6 shows the accuracy results of the ASR systems after testing the scientific speech audio files.

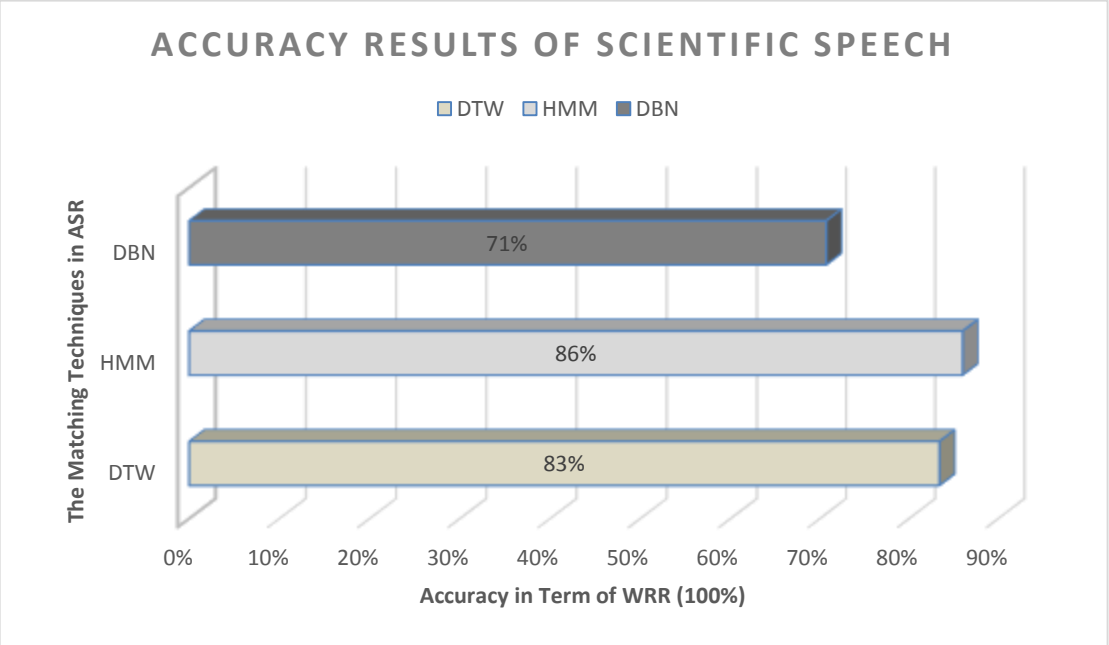


Figure 4.6. The accuracy results of the scientific speech audio files for several ASR combinations

Furthermore, a control category audio files were also tested in the ASR environments to extract the effects of matching techniques on the control speech. Thus, this category were added in the test samples in order to find the best matching technique that suits speech recognition applications which is the gate between human and machine interaction for doing human beings tasks. Table 4.8 shows the WER results of testing

control audio files in the ASR systems against DTW, HMM, and DBN matching techniques.

Table 4.8 The WER results of testing control category audio files in the environments of ASR methods

Controls Category Audio Files Results			
Audio File #	**DTW**	**HMM**	**DBN**
40	0	0	53.8
41	20	5.5	5.2
42	0	0	0
43	20	9.5	0
44	0	0	35
45	37.5	23.8	36.3
46	66.66667	38.8	0
47	12.5	12.9	13.3
48	25	23.8	0
49	0	0	35
50	75	41.3	15.38
Averages	**23.33333**	**14.14545**	**17.63455**

In order to represent the results in an understandable manner we provide a graph to show the differences in accuracy results of control category audio files. Figure 4.7 shows a chart that was used to represent the WER results of applying the control category audio files in the experiments.

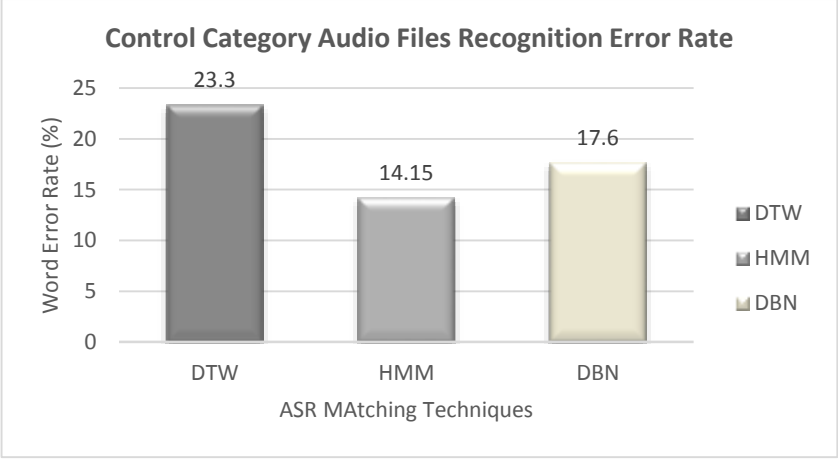


Figure 4.7. The WER averages chart after being tested in the ASR using several matching techniques

From control speech audio files point of view, the results showed that HMM achieved for the third time the best case with 14.15% in term of error rate result compared with DTW and DBN. In contrast, DTW went back to get the worst case with 23.3% which made with DBN matching technique. Intuitively, from accuracy perspective the HMM achieved 85.9% that registered as the most accurate result between them. As well as, the DTW achieved 76.7% which represent as the lowest accuracy result. Figure 4.8 shows the results of accuracy in term of WRR for the control speech audio files.

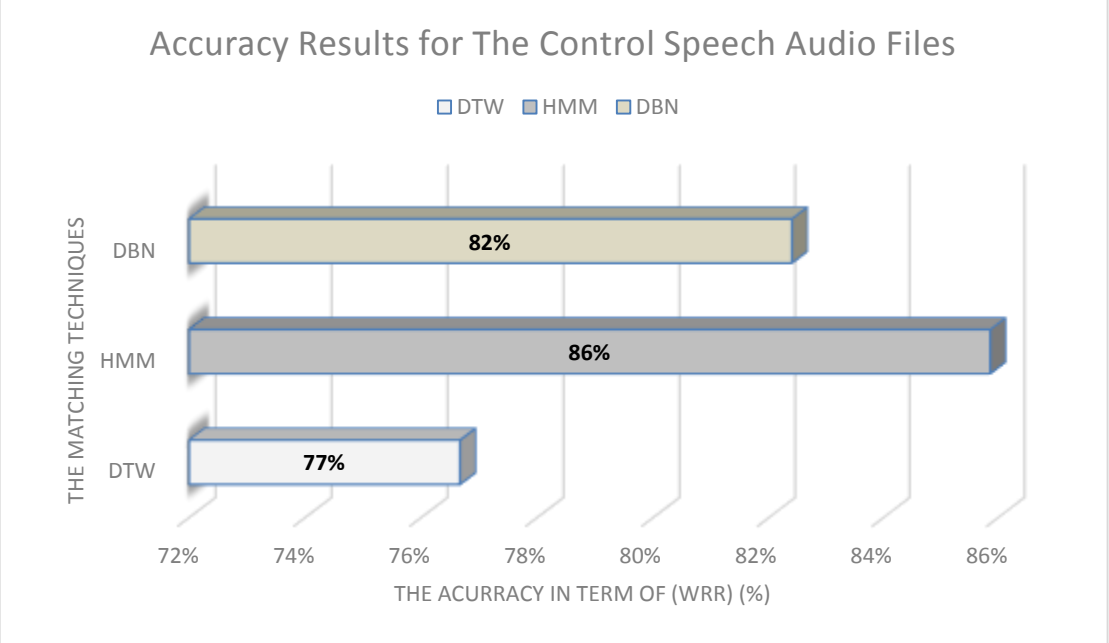


Figure 4.8. The accuracy results in term of WRR for control speech audio files were tested in ASR combinations

Consequently, we found that the accuracy results for each speech category was affected by the type of matching technique that was used in ASR system. Figure 4.9 shows a graph that represent the results of accuracy in term of WRR for each matching technique against speech file category.

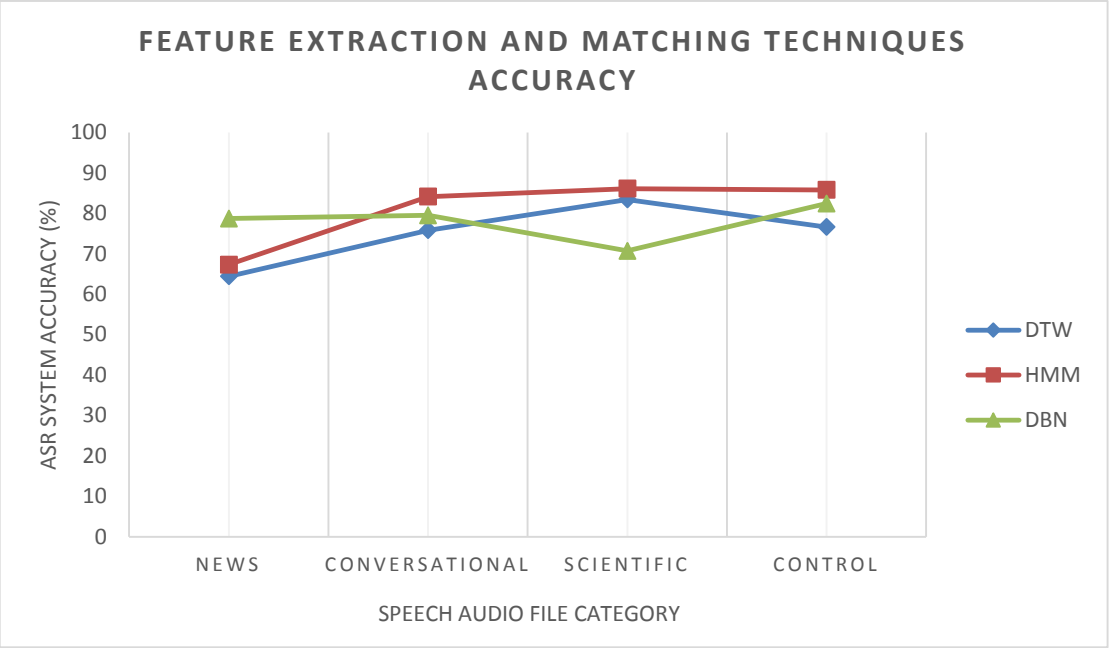


Figure 4.9. The accuracy graph of matching technique for each speech category in term of WRR

4.3. The experimental results of MT level

In this section, we described the examined experiments as well as we discussed the results of MT. For this purpose, the comparison that was discussed in chapter three took into consideration the translation model. Thus, the comparison was taken place between rule based MT, statistical based MT, and hybrid base MT. Therefore, we ran several experiments based on the results of ASR level that were discussed in section 4.2. Furthermore, the implementation of rule based translation model was conducted based on IBM translate engine, the implementation of statistical based translation model was conducted based on Google Translate API, and the implementation of hybrid based translation model was conducted based on SYSTRAN engine.

Consequently, the evaluation criteria which was taken into consideration in this level was by comparing the professional human manual translation with the MT results. For this purpose, we took WER also to find the error rate of online engines. Figure 4.10

shows an example of finding the WER of MT after applying the results of ASR for conversational category speech audio file against the translation models.

Original Speech Sentence:
"Would you like to go out for a meal in the evening?"
Manual Translation of Speech Audio:
هل تود الذهاب لتناول وجبة في المساء

Recognized Sentence: Using DTW
"Would you like to go out for a meal in the evening"

Translation using Google:
تريد أن تخرج لتناول وجبة في المساء؟
Translation Using IBM:
هل تريد الخروج للعشاء فلمسا اء
Translation Using SYSTRAN:
أردت أنت يحبّ أن ينصرف لوجبة في المساء؟

Recognized Sentence: Using HMM
"You like to go out for a meal in the evening"
Translation using Google:
كنت ترغب في الخروج لتناول وجبة في المساء
Translation Using IBM:
تريد أن أخرج للعشاء فى المساء
Translation Using SYSTRAN:
أنت تحبّ أن ينصرف لوجبة في المساء

Recognized Sentence: Using DBN
"Would you like to go out for a meal in this evening"
Translation using Google:
هل ترغب في الخروج لتناول وجبة في هذا المساء
Translation Using IBM:
هل ترغب في في الخروج لتناول وجبة في هذا المساء.
Translation Using SYSTRAN:
هل ترغب في الخروج لتناول وجبة في هذا المساء

	DTW	HMM	DBN
Google	11.41	12.5	20.4
IBM	19.3	27.5	25.53
SYSTRAN	20.9	18.18	18.6

Figure 4.10. Example of ASR and MT systems in Converting Speech files from English to Arabic

4.3.1. Google Translate experimental results

In this section, we discussed the results of Google Translate MT translation engine that represent the statistical based translation model. Therefore, the translation results of speech audio files found in appendices E and F. Figure 4.11 shows the Google Translate API webpage during extracting results of this study.

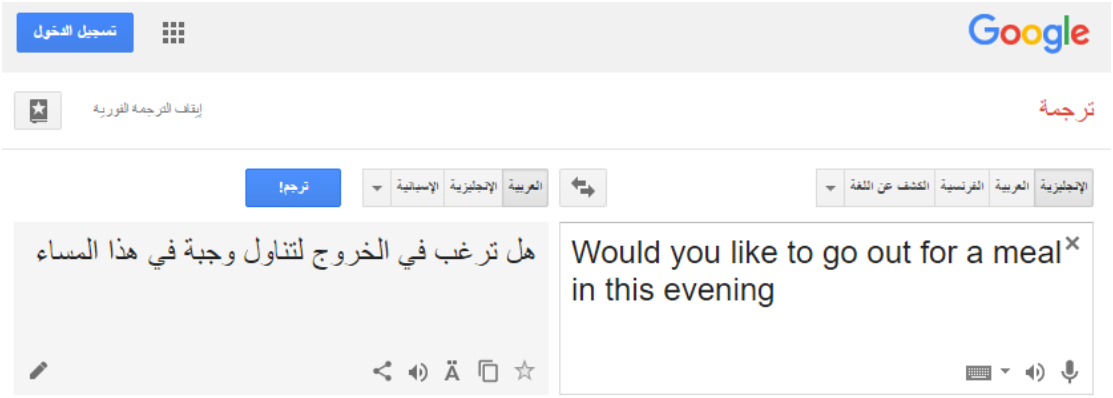


Figure 4.11. The online gate of Google Translate API engine

Table 4.9 shows the average WER for each speech category audio files translation using the recognized text from each matching technique.

Table 4.9. The average WER and the average of averages of MT based on the recognized sentences using ASR via DTW, HMM, and DBN of Google Translate API

	DTW	**HMM**	**DBN**
News	16.7	15.63	27.03
Conversational	19.9	7.89	51.4
Scientific Phrases	6.4	22.19	31.25
Controls	13.5	13.33	18.2
Average	**14.125**	**14.76**	**31.97**

From statistical translation based model, the results showed that the HMM and the DTW achieved the best error rate result compared with the DBN for news category. Furthermore, a drastic gap existed between the three classification techniques results in conversational category. Thus, the results showed a best case by using HMM in conversational category and a worst case by using the DBN technique. In the scientific phrase category the DTW showed best case with lower factor of WER compared with HMM. From control category perspective, the gaps in error rate were eliminated. Figure 3.12 shows the error rate results for classification techniques against the speech type for statistical based translation model.

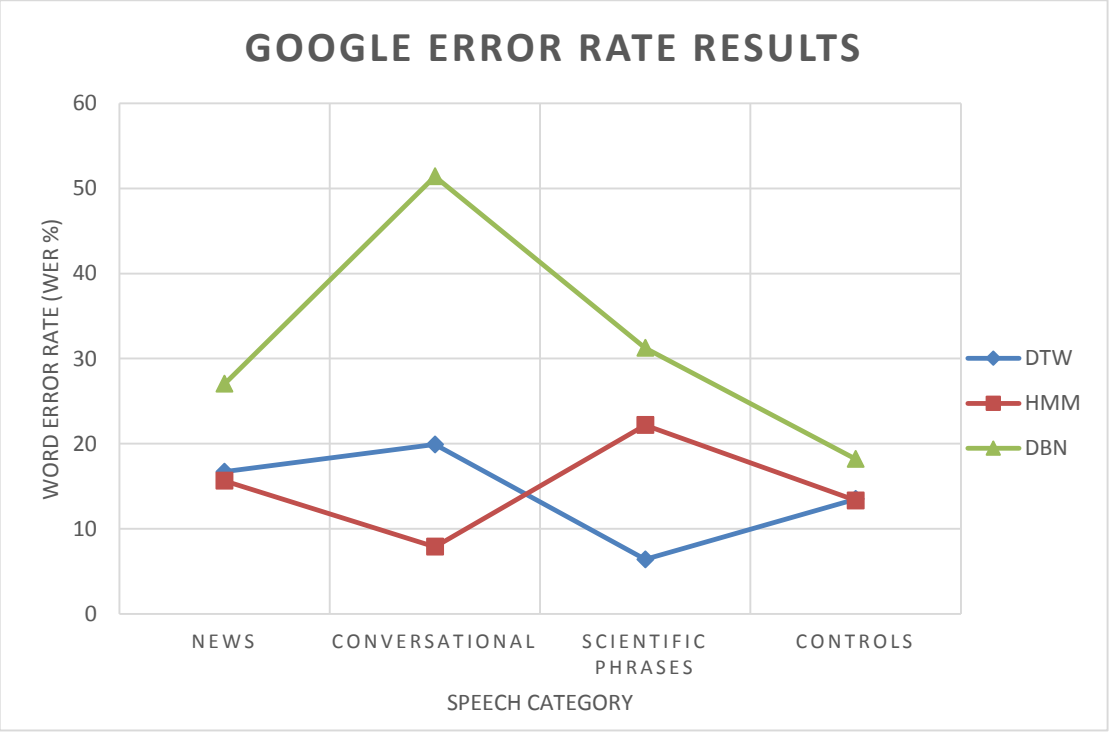


Figure 4.12. The average results of classification techniques in term of WER against each speech type in Google Translate API

4.3.2. IBM Translate experimental results

In this section, we discussed the results of IBM Translate engine which represents the rule based machine translation mode. Figure 4.13 shows the main gate of the IBM Translate API engine.

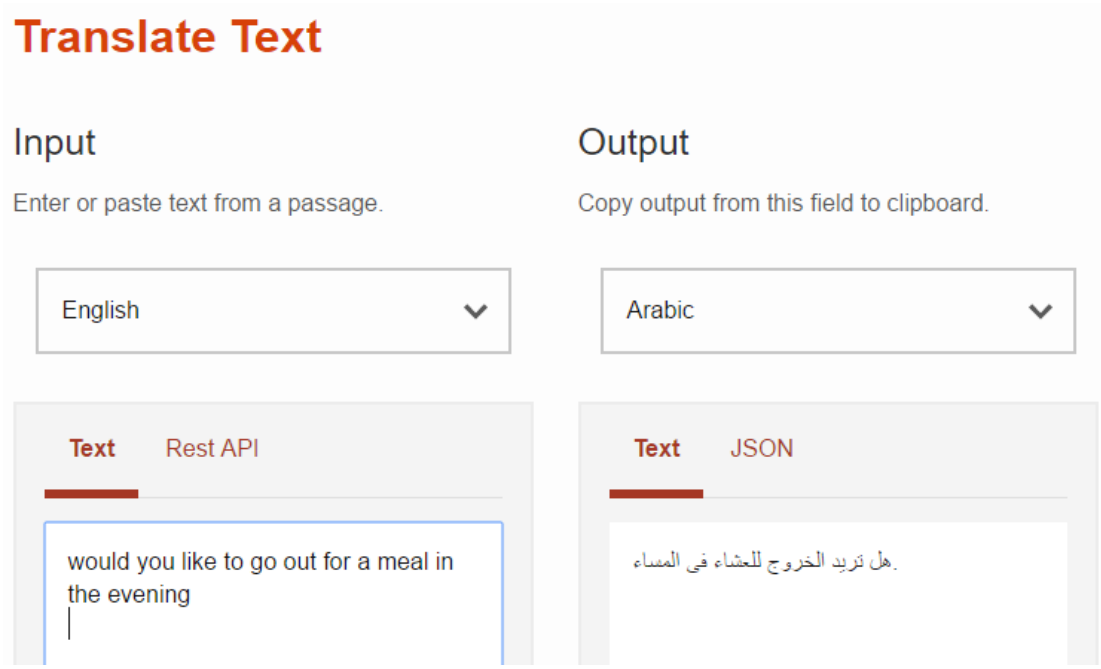


Figure 4.13. The online gate of IBM Watson Cloud Translate API engine

Table 4.10 shows the results of translating the recognized speech in term of the averages of WER.

Table 4.10. The average WER and the average of averages of MT based on the recognized sentences using ASR via DTW, HMM, and DBN of IBM Translate API

	DTW	**HMM**	**DBN**
News	9	1.6	11.6
Conversational	16.13	6.4	27.2
Scientific Phrases	10.71	13.5	18.3
Controls	2.5	8.03	14.52
Average	**13.93**	**7.38**	**17.91**

The results showed that IBM Watson Cloud which represents the rule based translation model had a several behaviors among speech audio file based on speech type and the used classification technique. Thus, the results showed that HMM in case of using the rule based translation model was the suitable classification technique for the news category as well as for the conversational one which is a positive effect. In contrast, The DBN classification technique showed a negative effect on both news and conversational speech categories. The error rate was balanced while using DTW with rule based translation model in news and conversational speech categories. Furthermore, DTW drawn a drastic gap in case of using it among rule based translation technique compared with the other techniques for both scientific phrases and controls categories which is a positive effect too. Figure 4.14 shows the error rate averages of using several ASR classification techniques.

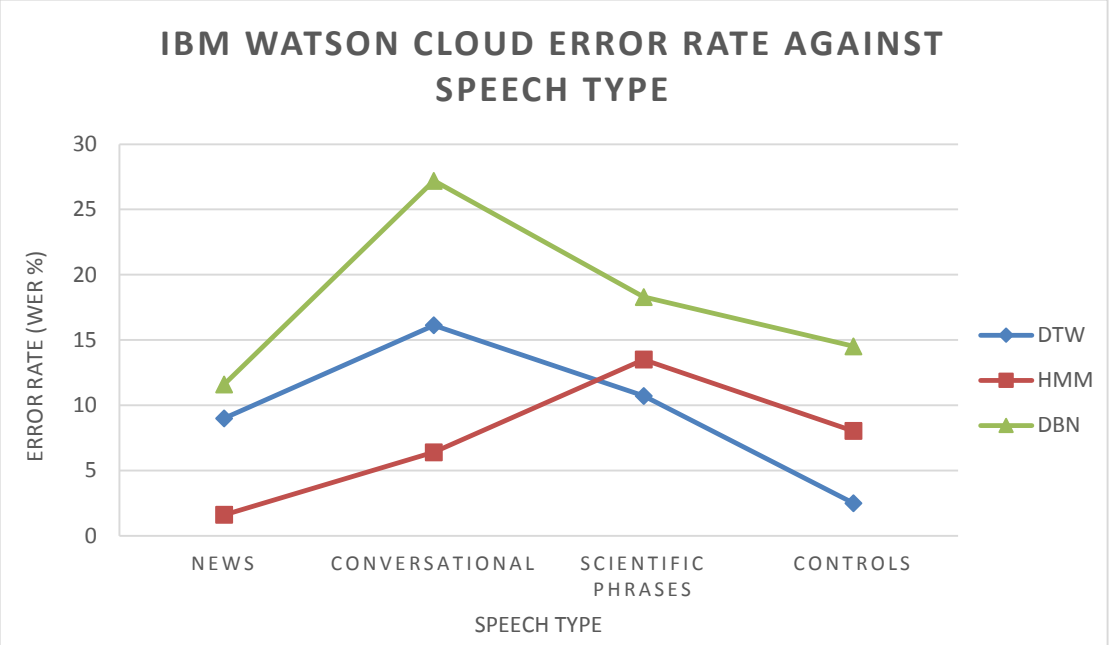


Figure 4.14. The average results of classification techniques in term of WER against each speech type In IBM Watson Cloud

4.3.3. SYSTRAN MT engine's experimental results

In this section, we discussed the results of translating the recognized speech audio file's sentences using the hybrid based machine translation that was provided by SYSTRAN MT engine. Figure 4.15 shows the main gate of SYSTRAN engine during sentence's translation phase.

Figure 4.15. The online gate SYSTRAN translation engine.

Table 4.10 shows the results of translating the recognized speech in term of the averages of WER.

Table 4.11. The average WER and the average of averages of MT based on the recognized sentences using ASR via DTW, HMM, and DBN of IBM Translate API

	DTW	**HMM**	**DBN**
News	32.9	21.63	43.75
Conversational	51.4	40.19	39.43
Scientific Phrases	31.25	32.24	53.87
Controls	18.2	10.32	36.2
Average	**33.43**	**26.095**	**43.31**

From hybrid based translation model perspective, the results showed a diversity in error rate results which reflects a diversity in translation accuracy by using the hybrid translation model. The HMM achieved the lowest error rate compared with the others in all speech types. The DTW and HMM had the same behavior in scientific phrases, as well as, the DBN and HMM had the same behavior in the conversational speech type. Figure 4.16 the error rate results for SYSTRAN engine among speech types after applying three ASR classification techniques.

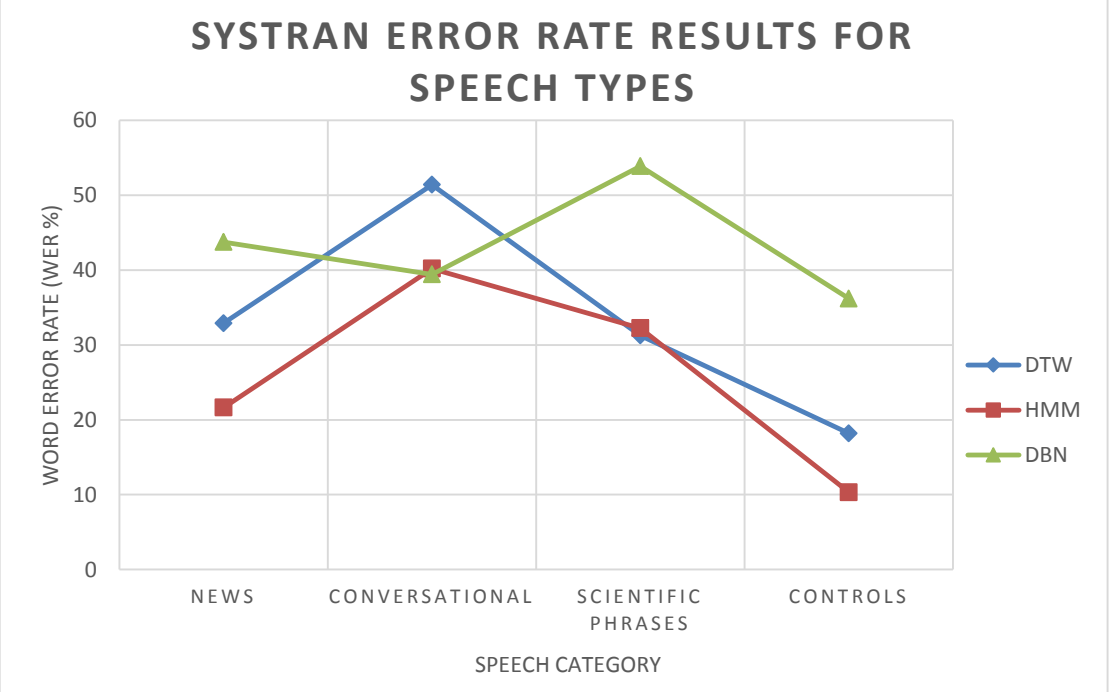


Figure 4.16. The average results of classification techniques in term of WER against each speech type In SYSTRAN engine

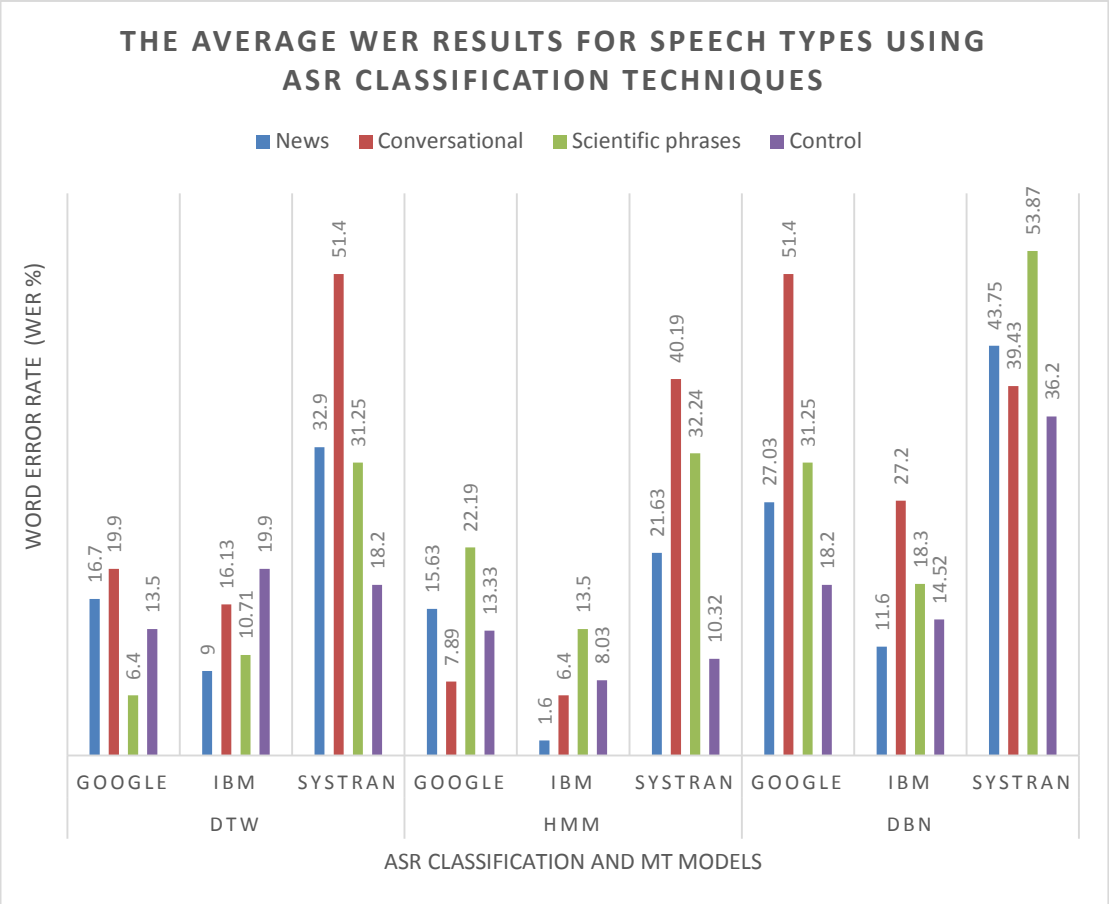


Figure 4.17. The average WER results for speech types using ASR classification techniques

The average values of WER results drew several gaps between the classification techniques and MT systems. For instance, In Statistical-based system using DTW classification technique the lowest WER achieved in case of scientific phrases, In contrast, this combination showed highest WER in conversational category. Thus, by choosing the HMM as a classification technique; the results showed that the lowest WER achieved in conversational category. Furthermore, by using the DBN as a classification technique a WER average results showed the lowest results can be achieved after applying the control category.

In rule-based system using DTW classification technique the lowest WER achieved in case of news category, In contrast, this combination showed highest WER in control category. Thus, by choosing the HMM as a classification technique; the results showed that the lowest WER achieved in news category. Furthermore, by using the DBN as a classification technique a WER average results showed the lowest results can be achieved after applying the news category. Consequently, these results showed by using the rule-based system in MT it had a positive effect in speech recognition and translation.

CHAPTER FIVE

Conclusions and Future Work

5.1. Overview

This chapter summarizes the conclusions of this study. Thus, the main theme of this study was to focus on measuring the effects of the classification techniques of the in ASR system on the MT translation models based on several types of speech file. Several experiments were conducted to measure the effects of matching approaches on the MT engine models, to show the effects of on the target language sentences (i.e. after translation), and to measure the accuracy in term of both WER and WRR for the target machine translation engines that has a speech recognition function for real time translation. Thus, this chapter was organized as section 5.2 to discuss the main conclusions, and section 5.3 to preview the future research works in the field of speech recognition and machine translation.

5.2 Conclusions

The main theme of this research was to measure the accuracy of ASR system classification technique with MT base system for real time speech translation from English to Arabic. In this study we ran several experiments to cover several environments and techniques to calculate the accuracy of the recognized and translated sentences in each combination. Nine environmental combinations were covered in the conducted experiments of feature extraction using MFCC with classification approaches (i.e. HMM, DTW, and DBN), and MT translation approaches (knowledge based approach, rule based approach, and hybrid based approach). For each combination we applied fifty speech audio files that covered four speech type categories (i.e. 14 sentences in news category,

15 sentences in conversational category, 10 sentences in scientific phrases category, and 11 sentences in control category). The empirical findings showed that the DBN as a classification technique achieved the best recognition rate with 79.2% compared with HMM and DTW for news category. However, the HMM classification technique achieved the best recognition rate for conversational with 80.1%, scientific phrases with 86%, and control with 63.8 % recognition rates. In contrast, using DTW as a classification technique in ASR had a negative behavior on the recognition rate for all speech categories.

On the other hand, the empirical findings from MT point of view the rule based model which was represented by IBM Watson cloud achieved the best results for in the majority of speech categories with 13.93% in conversational, 7.38% in scientific phrases, and 17.91% in control categories. The statistical based model – that was represented by Google Translate - in translation the empirical findings showed that for conversational and scientific phrases the error rate was close to rule based with an intangible difference. The hybrid translation based model influenced the error rate in the three ASR classification techniques and for all speech categories which was assigned as a negative effects.

5.3. Future Work

The empirical findings of this study was for fifty sentences that were segmented in word by word speech database. Thus, for future works increasing the number of training speech audio files is recommended to increase the accuracy of research findings as well as defining new speech categories is recommended too in order to specialize the English language for Arabic automatic translation systems. In the future, we recommend to work on combining two classification techniques such as DTW and HMM over a

several speech categories. As well as, we recommend to increase the number of words in each sentences to cover continues speech for more than twenty seconds. Furthermore, we recommend to work on optimizing the results of MT by using the hybrid-based rules depend on the type of speech. The Arabic language contains a lot of grammar rules. Therefore, it is recommended to enhance the data MT systems with a grammar dictionary in order to be used in a WEB-API's.

References

Abdel-Hamid, O., Mohamed, A., Jiang, H., Peng, L., Penn, G., and Yu, D. (2014). Conventional Neural Network for Speech Recognition. *ACM Transaction on Audio Speech, and Language Processing, 22 (10), PP. 2329-2339.*

Alotaibi, Y., and Hussein, A. (2010). Comparative Analysis of Arabic Vowels Using Format and an Automatic Speech Recognition System. *International Journal of signal processing, image processing and pattern recognition, 3(2), PP. 11-22.*

Alsuliaman, M., Muhammad, G., Bencherief, M., Mahmooud, A., Ali, Z, and Al-Jabri M. (2011). Building Rich Arabic Speech Database. *IEEE Fifth Asia Modeling Symposium, 3 (1), PP. 100-105.*

Antony, J. (2013). Machine Translation Approaches and Survey for Indian Languages. *Computational Linguistics and Chinese Language Processing, 18 (1), PP. 47-78.*

Baker, J., Deng L., Glass, J., Khudanpur, S., Chin-hui L., Morgan, N., and O'Shaughnessy, D. (2009). Developments and Directions in Speech Recognition and Understanding, *Signal Processing Magazine, IEEE, 26 (3), PP.75-80, May 2009*

Benzeguiba, M., Mori, R.D., Deroo, O., Dupon, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., Automatic Speech Recognition and Speech Variability: a Review, *Speech Communication (2007).*

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic Speech Recognition for Under Resourced Languages: A Survey. *Speech Communication, 56 (1), PP. 85-100.*

Chapaneri, S. (2012). Spoken Digits Recognition Using Weighted MFCC and Improved Feature for Dynamic Time Wrapping. *International Journal of Computer Applications. 4 (3), PP. 6-12*.

Cutajar, M., Gatt, E., Grech, I., Casha, O., and Micallef, J. (2013). Comparative Study of Automatic Speech Recognition Techniques. *The Institution of Engineering and Technology, 7 (1), PP. 25-46*.

Essa, E., Tolba, A., and Elmougy, S. (2008). Combined Classifier Based Arabic Speech Recognition. *International Journal in Speech Recognition and Computer-Human Interaction. 4 (2), PP. 11-15*.

Franklen, J., West, M., and King, S. (2007). Articular Feature Recognition Dynamic Bayesian Network (DBN).*Computer Speech and Language Conference, PP. 35-70.*

Garg, A. and Rehg, V. (2011). Audio-Visual Speaker Detection Using Dynamic Bayesian Network. *The Institution of Engineering and Technology 1 (1), PP. 19-27*.

Gemmeke, J., Virtanen, T., and Demuynck, K. (2013). Exemplar-Based Joint Channel and Noise Compensation. *IEEE International Conference on Acoustic, Speech and Signal Processing.*

Ghahramani, Z. (2001). An Introduction to Hidden Markov Models and Bayesian Networks. *International Journal of Pattern Recognition and Artificial Intelligence, 15 (1), PP. 9-42*.

Giannoluious, P., and Patamins, G. (2012). A Hierarchical Approach with Feature Selection for Emotion Recognition From Speech. *Proceeding of the Eight International Conference on Language Resources and Evaluation LREC – 2012, Istanbul, Turkey, 1203-1206. ISBN: 978-951-17408-7-7.*

Hammo, B., Sleit, A., El-Haj, M, Baarah, A., and Abu-Salem, H., (2012). A Computational Approach for Identifying Quranic Theme. *International Journal of Computing Proceeding Oriental Language. 22 (4), 189-196*.

Jouvet, D., and Vinusea, N. (2012). Classification Margin for Improved Class Based Speech Recognition Performance. *IEEE International Conference on Acoustic, Speech and Signal Processing, Kyoto, Japan, PP. 4285-4288.*

Kazuma Nishimura, Hiromichi Kawanami, Hiroshi Saruwatari and Kiyohiro Shikan (2011). Investigation of Statistical Machine Translation Applied to Answer Generation for a Speech-Oriented Guidance System. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, PP. (11-18).*

Kurzekar, P., Deshmukh, R., Waghmare, V., and Shrishrimal, P. (2014). A Comparative Study of Feature Extraction Techniques for Speech Recognition System. *International Journal of Innovative Research in Science, Engineering and Technology, 3 (12), PP. 18006-180016.*

Lahdesmaki, H., and Shumleuch, A. (2008). Learning the Structure of Dynamic Bayesian Networks from Time Series and Steady state Measurements. *Machine Learning (ML), 71 (2), PP. 185-217.*

Lei, X., Senior, A., Gruenstein, A., and Sorensen, J. (2013). Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices. *14th Annual Conference of International Speech Communication Association, PP. 662-665*.

Livescu, K, Bilmes, J., and Glass, J. (2003). Hidden Feature Model for Speech Recognition Using Dynamic Bayesian Networks. *8th European Conference on Speech Communication and Technology*.

Livescu, K., Glass, J., and Bilmes, J. (2013) Hidden Feature Model for Speech Recognition Using Dynamic Bayesian Networks. *8th European Conference on Speech Communication and Technology*

Mamta, A. and Wala, T. (2015). A Review of Various Approaches for Machine Translation. *International Journal of Advance Research in Computer Science and Management Studies, 3 (2), PP. 108-113*. ISSN: 2321-7782.

Muda, L., Begam, M., and Elamvazuthi, I. (2010). *Voice Recognition Algorithm Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Wrapping (DTW) techniques*. Journal of computing, 2 (3), PP. (138-143). ISSN: 2151-9917.

Ng., R., Shah, K., Aziz, W., Specia, L., and Hain, T. (2015). *Quality Estimation for ASR K-Best List Rescoring in Spoken Language Translation*. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, 5226-5230, ISBN: 978-4677-6997-8.

Och, F. and Ney, H. (2002). Discriminative and Maximum Entropy Models for Statistical Machine Translation. *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), PP. 295-302*

Park, Y., Patwardhan, S., Visweswariah, k., and Gates, S. (2008). An Empirical Analysis of Word Error Rate and Keyword Error Rate. *Proceeding of the International Conference on Spoken Language Processing, PP. 2070-2073.*

Paul, D. (1990). Speech Recognition Using Hidden Markov Model. *The Lincoln Laboratory Journal- Journal of Computer Science. 3 (1), PP. 41-62.*

Rabiner, L, Schafer, R. (2014). MATLAB Exercises in Support of Teaching Digital Speech Processing. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, PP. (2480-2483).*

Rajhona, J. and Pollak, P., (2011). ASR System in Noisy Environment Analysis and Solution for Increasing Noise Robustness. *Radio Engineering, 20 (2), PP. 74-84.*

Rhaman, K. and Tarannum, N. (2012). A Rule Based Approach for Implementation of Bangla to English Translation. *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT).*

Sharwanker, V and Thakare, V. (2013). Techniques for Feature Extraction in Speech Recognition System: A Comparative Study. *Computer and Information Sciences, 6 (1), 58 – 69. ISSN: 1913-8989.*

Singh, N., Khan, R. A., & Shree, R. (2012). MFCC and Prosodic Feature Extraction Techniques: A Comparative Study. *International Journal of Computer Applications IJCA, 54 (1), PP. 9-13.*

Stephenson, T., Bourland, H., Bengio, S., and Morri, A. (2000). Automatic Speech Recognition Using Dynamic Bayesian Networks with both Acoustic and Articular Variables. *6th International Conference on Spoken Language Processing (ICSIP'00) China, PP. 951-954.*

Syahrina, A. and Lind, B. (2011). Online Machine Translator System and Result Comparison. *University of Boras, School of Computing, 1 (3), PP. (18-26).*

Vimala, C. and Radha, V. (2012). A Review on Speech Recognition Challenges and Approaches. *World of Computer Science and Information Technology Journal (WCSIT), 2 (1), PP. (1-7). 2221-0741*

Watanbe, S., and LeRoux, J. (2014). Black Box Optimization for Automatic Speech Recognition. *IEEE International Conference on Acoustic Speech and Signal Processing ICASSP-2014, Fortena, Italy.* PP. 3256-3260.

Yadav, K. and Mukhedkar, M. (2013). Review on Speech Recognition. *International Journal of Science and Engineering, 1 (2), PP. 61-70*. ISSN: 2347.

Zhang, Y., Adl, K., and Glass, J. (2014). Fast Spoken Query Detection Using Lower Bownd Dynamic Time Wrapping on Graphical Processing. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, PP. 5173-5176.*

Zhu, S. and Wang, Y. (2015). Hidden Markov Induced Dynamic Bayesian Network for Recovering Time Evolving Gene Regulatory Networks. *Scientific Reports, 1 (4), PP. (1-17).*

Mon, S. and Tun, H. (2015). Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM). *International Journal of Scientific and Technology Research (IJSTR), 4 (6), PP. 349-352.*

Mishra, K., Bhagat, P., and Kazi, A. (2016). Automatic Subtitle Generation for Sound in Videos. International Journal of Engineering and Technology (IRJET)

3 (2), PP. 915-918.

Appendices

Appendix A

Table shows the sentences of speech files for the conversational category.

Number	The Test Sentence	Category
1.	Hello world	Conversational
2.	what are you doing now	Conversational
3.	software bugs leads for system down	Conversational
4.	Would you like to go out for a meal in the evening?	Conversational
5.	Please make up clean the room	Conversational
6.	Do you have set meals menu	Conversational
7.	The traffic light is red	Conversational
8.	We want to have a table near the window	Conversational
9.	the computing power changed the life style	Conversational
10.	Do you like the middle east university?	Conversational
11.	The boys are watching television right now	Conversational
12.	my wedding will be in the next Friday	Conversational
13.	I break the routine and I find new exercises	Conversational
14.	Thank you for coming	Conversational
15.	find the suitable book from library	Conversational

Appendix B

Table shows the sentences of speech files for the scientific phrases category.

Number	The Test Sentence	Category
1.	Working on speech recognition system	Scientific Phrases
2.	developing new applications using cloud computing	Scientific Phrases
3.	system development life cycle in project management	Scientific Phrases
4.	the top ten myths on e-commerce	Scientific Phrases
5.	paravirtualization hypervisor using Xen	Scientific Phrases
6.	foreign key in database management system	Scientific Phrases
7.	computer information system enhanced application	Scientific Phrases
8.	agile methodology versus waterfall methodology	Scientific Phrases
9.	programming languages and compilers	Scientific Phrases
10.	dynamic programming in speech recognition	Scientific Phrases

Appendix C

Table shows the sentences of speech files for the control phrases category.

Number	The Test Sentence	Category
1.	write a message	Control Phrases
2.	open my mail inbox	Control Phrases
3.	find my files	Control Phrases
4.	recognize the speech	Control Phrases
5.	check the system performance	Control Phrases
6.	view files information	Control Phrases
7.	send these files via Bluetooth	Control Phrases
8.	view computer system information	Control Phrases
9.	print the hidden files	Control Phrases
10.	search for machine translation in google	Control Phrases
11.	delete recent files from downloads folder	Control Phrases

Appendix D

Table 1 shows the results of speech recognition using MFCC with DTW technique of news category.

Results of MFFC + DTW
Things you can learn from their history.
Do you do anything used embarrassing things.
He didn't want to quit or Sunday.
It is just breaking my heart.
Where you can go in and who's gonna.
Welcome to the new server or.
Did Charles who won the hearts of four.
Suicide of Hanukkah.
How to publish a book.
Are you trying to tie the two party.
Discover the first tune that lies hidden in your salad.
Do you make these mistakes in English.
However improved my memory in one evening.
How to win friends and influence people

Table 2 shows the results of speech recognition using MFCC with DTW technique of conversational category.

Hello world.
What are you doing now
Software bugs leads for system down.
Would you like to go out for a meal in the evening
Please may come in the room.
You have said you know.
The traffic light is red.
We want to have a table near the window.
The computing power to the lifestyle.
Yeah like the Middle East university.
You boys aren't watching television right now
My wedding will be the next Friday.
Alright breaks they don't you know I'm fairly new exercises.
Thank you for coming.
Find this user will book from the library.

Table 3 shows the results of speech recognition using MFCC with DTW technique of Scientific phrases category.

Walking in a speech recognition system.
Developing new applications using cloud computing.
System development lifecycle project management.
The top ten myths ecommerce.
Realization I wasn't using Zen.
Forming a key in database management system.
Computer information systems enhanced applications.
I'm John methodology verses watchful methodology.
Programming languages and compilers.
Dynamic programming speech recognition.

Table 4 shows the results of speech recognition using MFCC with DTW technique of control category.

Write a message.
Open my main inbox.
Find my files.
Recognize this speech.
Check the system performance.
You files information.
Send these files whatever you choose.
You computer system information.
Then the hidden files.
Search for machine translation in Google.
Do you need to listen files from downloads for

Appendix E

Table shows the results of speech recognition of speech categories using MFCC and HMM technique

News Category	Things you can learn from the apple store.
	Do you do I leave these denim barcy things?
	Who want to quit work Sunday?
	my facebook is making life hard
	I'm doing a good used car.
	Welcome to the new civil war.
	Charge who won the hearts of all.
	She sighed heika.
	How to publish a book?
	I'm tired of the party.
	Discover the first one that lies ago did you send it.
	Do you make this mistake in English?
	How I put my memory in one evening?
	how to win friends and influence people
Conversational Category	hello world
	What are you doing now?
	Software bitlis for system down.
	You like to go out for a meal in the evening.
	Please make up clean the room.
	Do you have said communion?
	The traffic light is red.
	Do you want to have a table near the window?
	New computing power change the last bite.
	Do you like any of these universities?
	The boys are watching television right now.
	Why would they will be in the next Friday?
	Hi Bridget courteen on finding you some slices.
	thank you for coming
	Find a suitable book from the library.

Scientific Phrases	Working on the speech recognition system.
	Open your vacation using cloud computing.
	System development life cycle in project management.
	Top 10 minutes only commers.
	Visualizacion have a good news and then.
	Put ink in database management system.
	Computer information systems in homes application.
	Hi John this is gonna do you guys want to forward his own.
	Programming languages are compilers.
	Dynamic programming in speech recognition.
Control Category	Why is the message?
	Open my email inbox.
	Find my fine.
	We could night is this speech.
	Check the system performance.
	Do you find information?
	These files via Bluetooth.
	If you computer system information?
	Plant they have the files.
	Search for machine translation Google.
	Delete recent files from downloads folder.

Appendix F

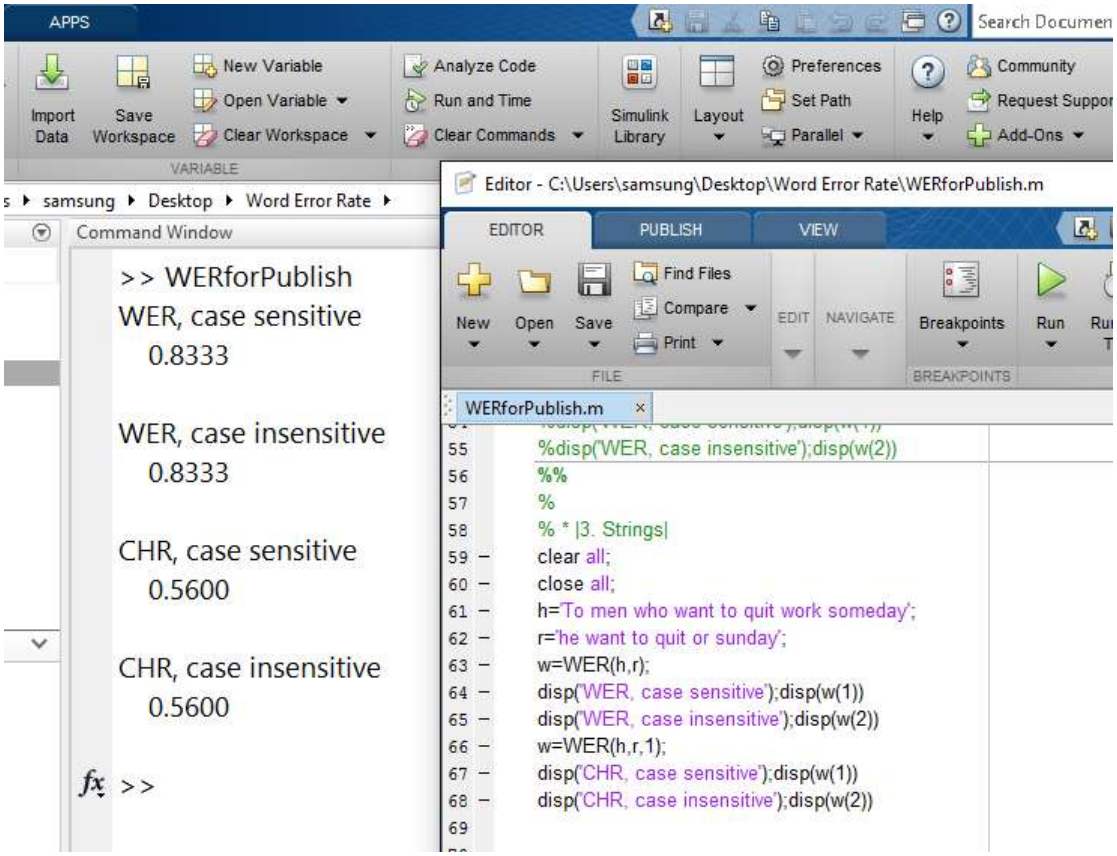
Table shows the results of speech recognition of the speech categories using the MFCC and DBN

News Category	Things you can learn from the Apple Store
	do you do any of these 10 embarrassing things
	two men who want to quit work someday
	my Facebook is making life hard
	where you can get good used car
	will turn to the u.s. Civil War
	the charge who won the hearts of all
	suicide of
	how to publish a book
	all you ever think that I'd party
	discovered the first one that lies hidden in your salary
	you make these mistakes in English
	how do I improve my memory in one evening
	how to win friends and influence people
Conversational Category	out of work
	what are you doing now
	so your bus leaves for system down
	would you like to go out for a meal in this evening
	please make up in the room
	you have hit me up when you get
	The traffic light is red.
	Do you want to have a table near the window?
	the computing power changed the life style
	do you like the middle east university?
	it was all watching television right now
	my good thing won't be in the next Friday
	off breaks his routine and I find nutrisciences
	thank you for coming
	for you to come home for from Library

Scientific phrases Category	working in speech recognition system
	developing in your kitchen using cloud computing
	system look like in project management
	ten myths ecommerce
	visulization I wasn't using Zen.
	forigen key in database management system
	Computer information system enhanced application
	a job with bulging vs. waterfall methodology
	Programming languages are compilers.
	dynamic programming in speech
Control Category	to Michelle Parker message
	open my email inbox
	find my files
	recognize the speech
	to the speech to the system performance
	so far as information
	send these files via Bluetooth
	to computer system information
	print the hidden files
	president 561 machine translation Google
	read recent files from download folder

Appendix G

The WER measuring MATLAB program of a test sentence ''*To men who want to quit work someday*' and the result sentence '*he want to quit or sunday*'. The WER tool box shows the WER of case sensitive and without case sensitive.



Appendix H

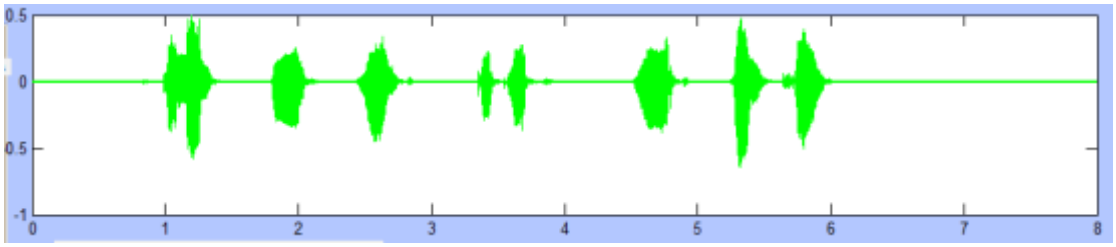
Table shows the results of WER in the level of speech recognition in each speech category.

	WER (Original Vs. DTW)	WER (original Vs. HMM)	WER (Original Vs. DBN)
News Category	42.85714286	33.33	0
	42.85714286	84.4	11.11111
	71.42857143	45.5	12.5
	42.85714286	75.2	16.66667
	50	34.2	25
	33.33333333	25	80
	50	28.9	12.5
	33.33333333	36.8	33.33333
	0	9	0
	71.42857143	45.7	42.85714
	37.5	11.1	36.36364
	0	5.3	14.28571
	22.22222222	11.9	12.5
	0	10.6	0
Conversational Category	0	0	63.4
	0	0	0
	0	0	23.5
	0	2	3.9
	66.66666667	33.3	15.3
	88.88888889	68.1	54.8
	0	0	0
	0	0	13.9
	28.57142857	22.2	0
	55.55555556	16.7	0
	28.57142857	11.3	21.9
	0	8.9	22.5
	43.8	42.11	36.17
	0	0	0
	50	34.15	51.3
	0	0	2.7

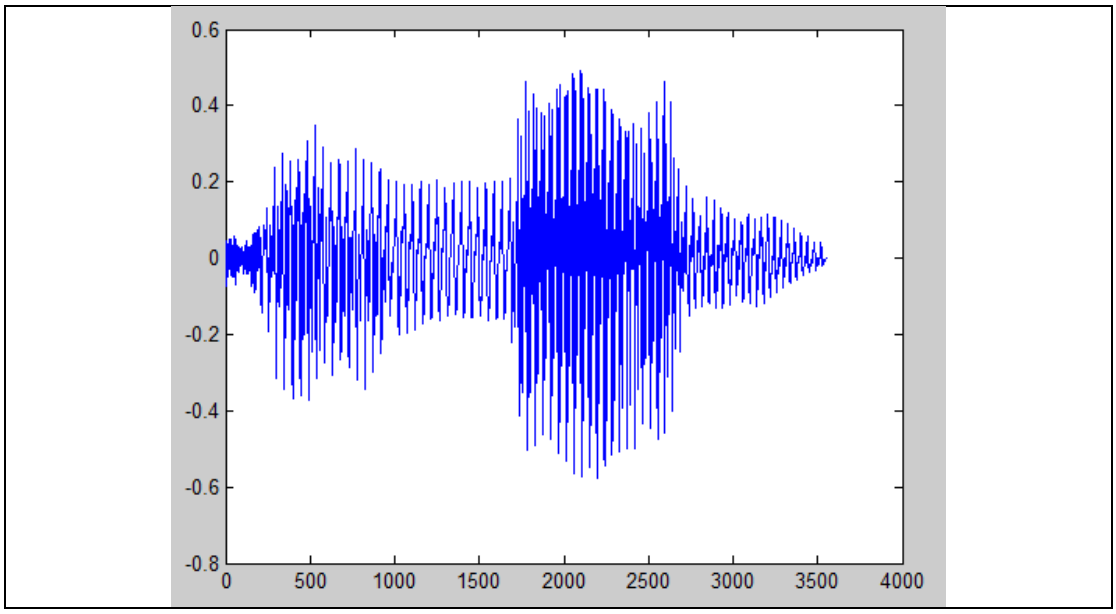
Scientific Phrases	**42.5**	**0**	**29.17**
	8.5	**8.5**	**44.7**
	12.9	**14.8**	**63.1**
	62.07	**58.9**	**56.6**
	9.3	**11.63**	**0**
	2.3	**4**	**0**
	25.5	**78.6**	**40**
	0	**2.86**	**8.5**
	2.8	**10.5**	**48.2**
Control Category	**0**	**0**	**53.8**
	20	**5.5**	**5.2**
	0	**0**	**0**
	20	**9.5**	**0**
	0	**0**	**35**
	37.5	**23.8**	**36.3**
	66.66666667	**38.8**	**0**
	12.5	**12.9**	**13.3**
	25	**23.8**	**0**
	0	**0**	**35**
	75	**41.3**	**15.38**

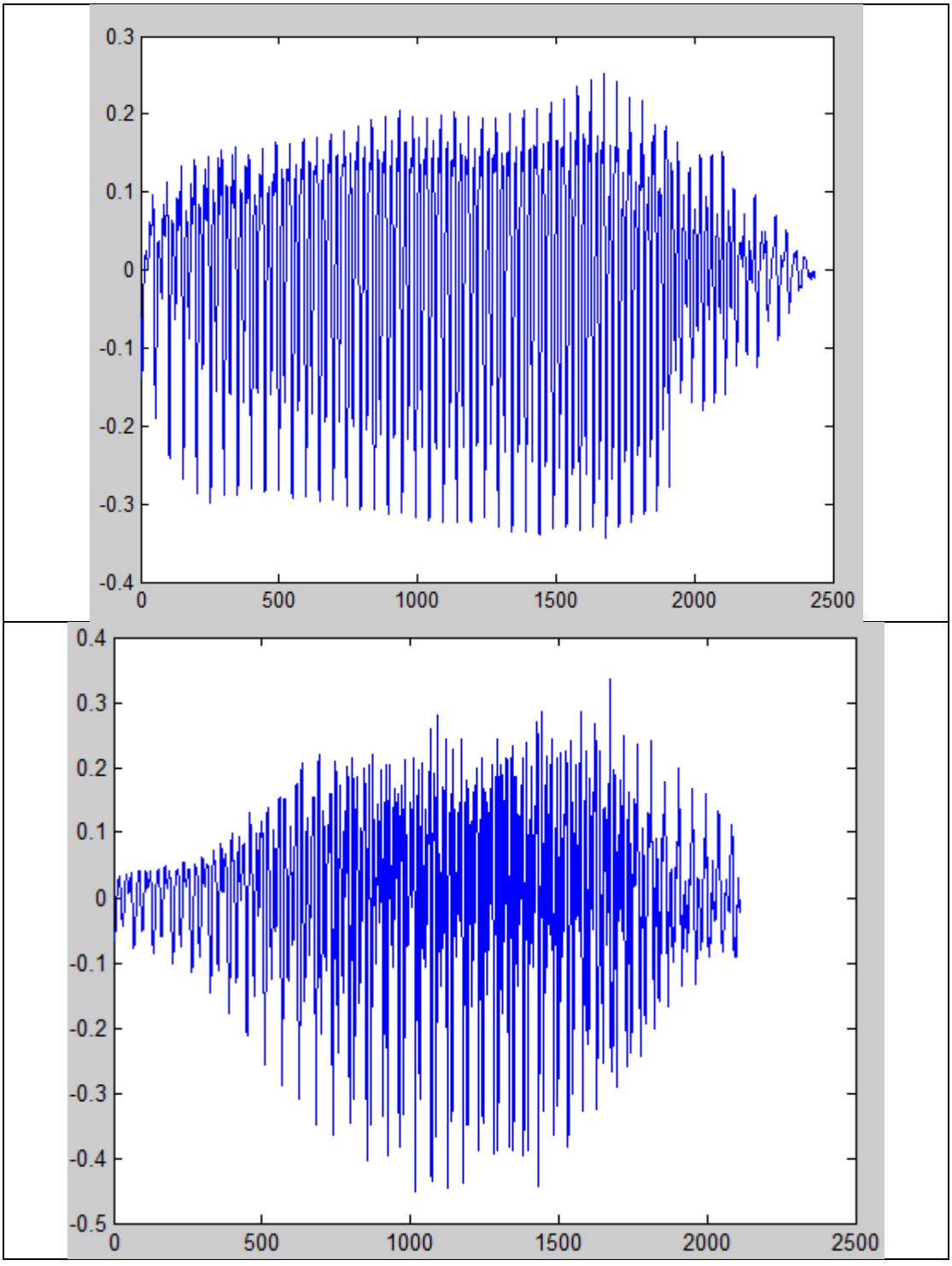
Appendix I

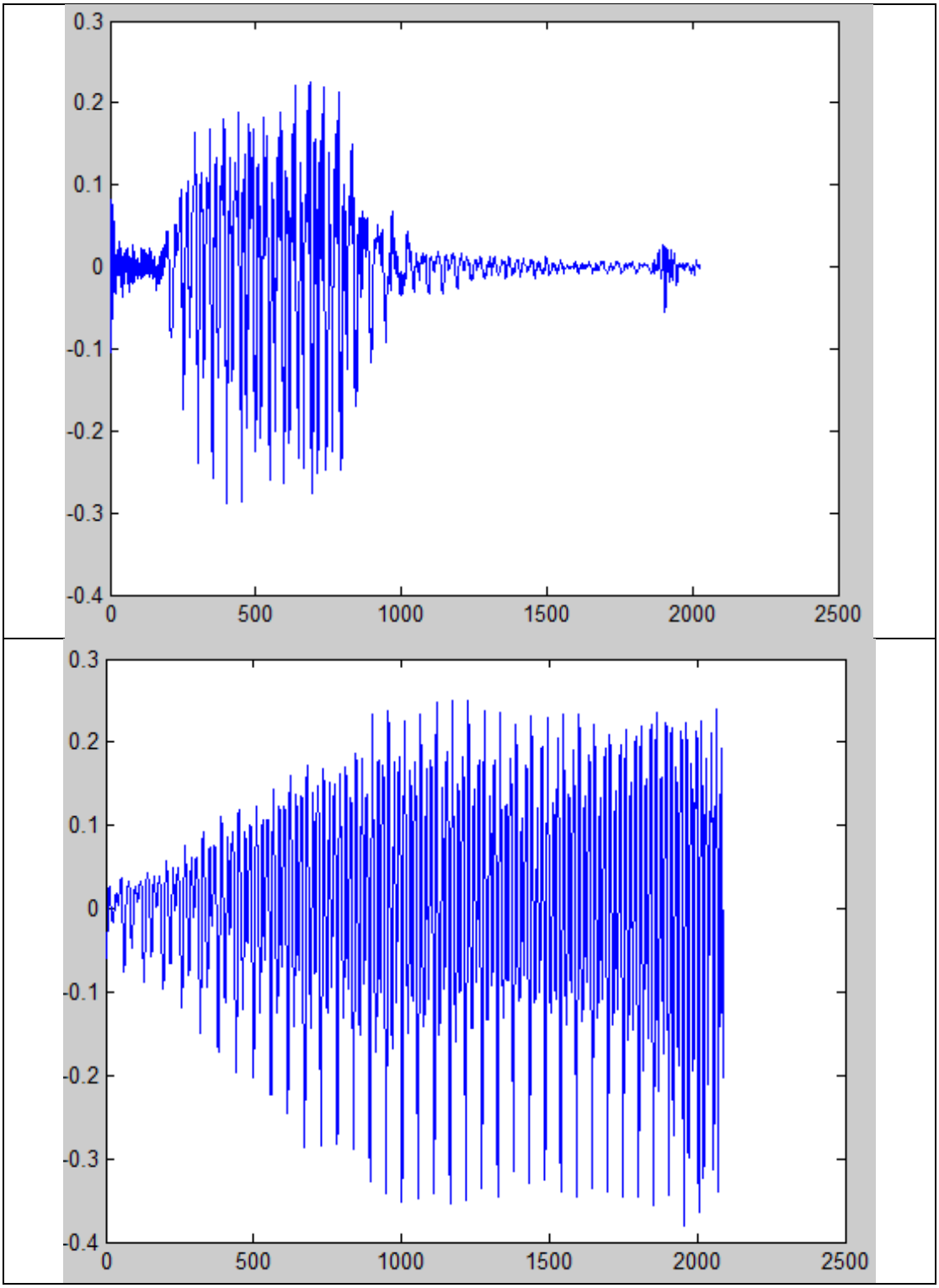
Speech signal of test sentence 'To men who want to quit work someday'.

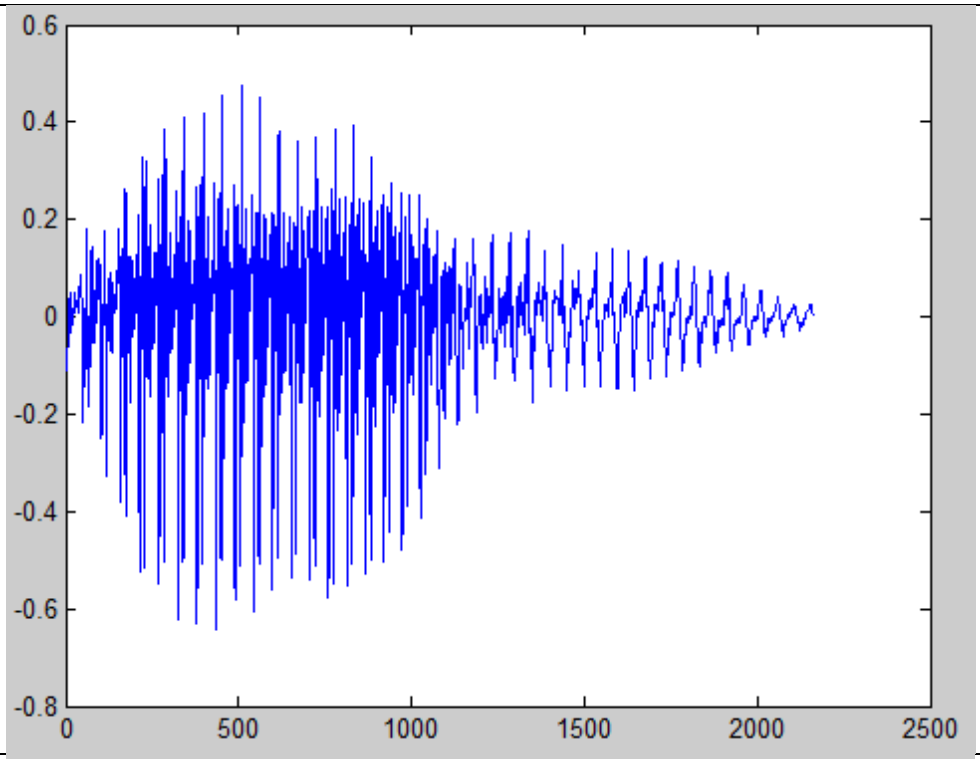


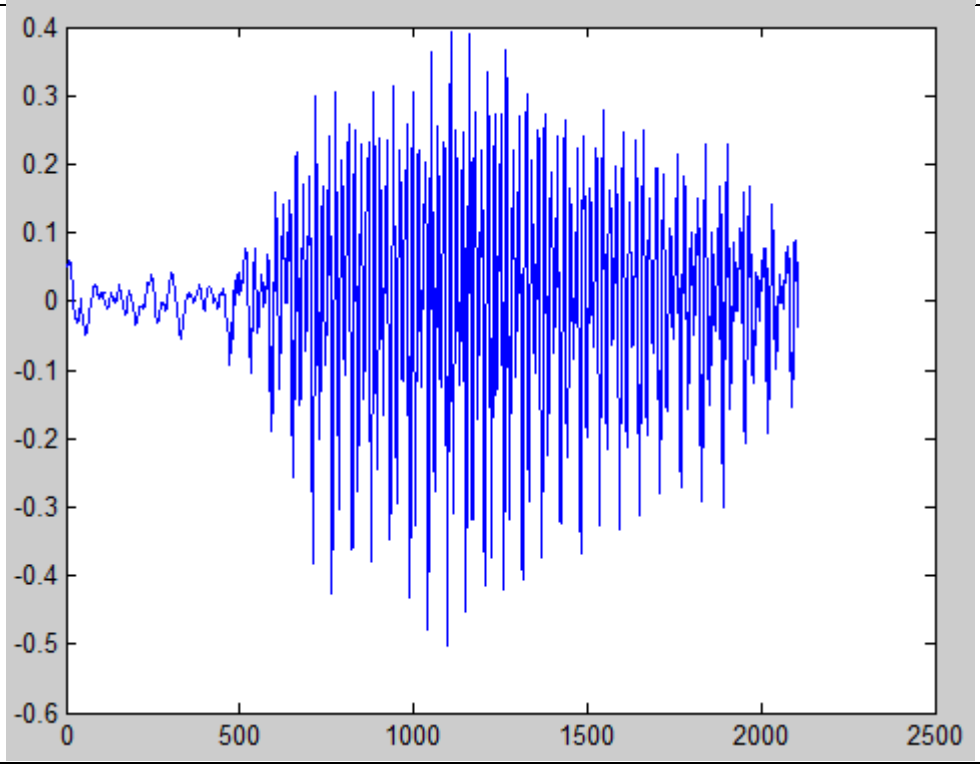
Speech segmentation signals for each word {'To men', 'whom', 'want', 'to', 'quit', 'work', 'someday'}.




0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
500
1000
1500
2000
2500
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
0
500
1000
1500
2000
2500

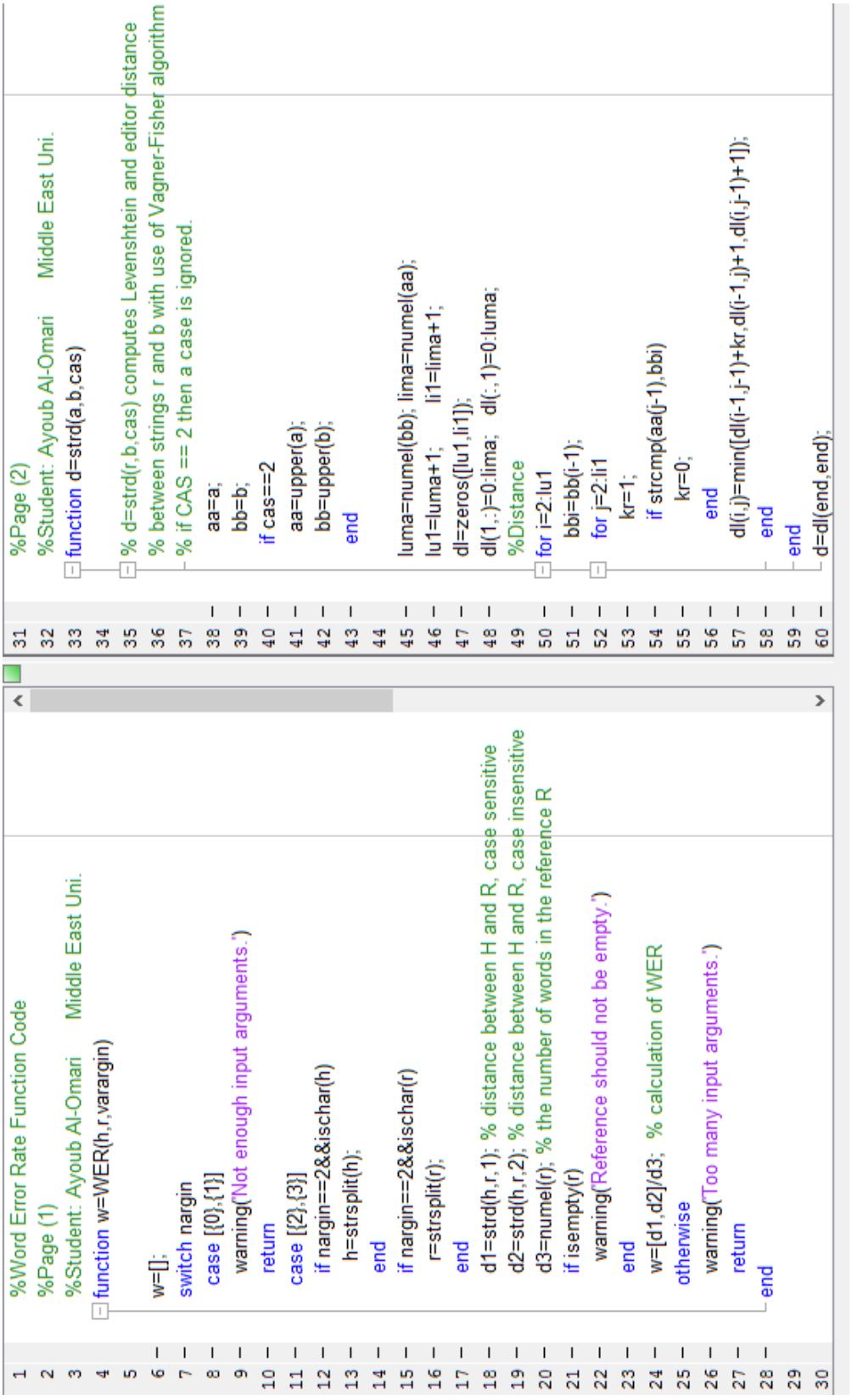

0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
500
1000
1500
2000
2500
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
500
1000
1500
2000
2500


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
500
1000
1500
2000
2500


0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
0
500
1000
1500
2000
2500

Appendix J

The MATLAB code of WER



```
%Word Error Rate Function Code
%Page (1)
%Student: Ayoub Al-Omari    Middle East Uni.
function w=WER(h,r,varargin)
    w=[];
    switch nargin
     case [{0},{1}]
      warning('Not enough input arguments.')
      return
     case [{2},{3}]
      if nargin==2&&ischar(h)
        h=strsplit(h);
      end
      if nargin==2&&ischar(r)
        r=strsplit(r);
      end
      d1=strd(h,r,1); % distance between H and R, case sensitive
      d2=strd(h,r,2); % distance between H and R, case insensitive
      d3=numel(r); % the number of words in the reference R
      if isempty(r)
        warning('Reference should not be empty.')
      end
      w=[d1,d2]/d3;  % calculation of WER
     otherwise
      warning('Too many input arguments.')
      return
    end
```

```
%Page (2)
%Student: Ayoub Al-Omari    Middle East Uni.
function d=strd(a,b,cas)
% d=strd(r,b,cas) computes Levenshtein and editor distance
% between strings r and b with use of Vagner-Fisher algorithm
% if CAS == 2 then a case is ignored.
    aa=a;
    bb=b;
    if cas==2
     aa=upper(a);
     bb=upper(b);
    end

    luma=numel(bb); lima=numel(aa);
    lu1=luma+1;     li1=lima+1;
    dl=zeros([lu1,li1]);
    dl(1,:)=0:lima;   dl(:,1)=0:luma;
    %Distance
    for i=2:lu1
     bbi=bb(i-1);
     for j=2:li1
       kr=1;
       if strcmp(aa(j-1),bbi)
         kr=0;
       end
     dl(i,j)=min([dl(i-1,j-1)+kr,dl(i-1,j)+1,dl(i,j-1)+1]);
     end
    end
    d=dl(end,end);
```