إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Adaptation of Acoustic and Language Model for Improving Arabic Automatic Speech Recognition

مؤائمة النموذج الصوتي واللغوي لزيادة فاعلية التعرف التلقائي على المنطوق العربي

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وإن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

DECLARATION

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification.

Student's name: اسم الطالب: **أسامة سليمان إنشاصي**

Signature: التوقيع:

Date: التاريخ: 2016/2/1 م

بسم الله الرحمن الرحيم

Islamic University – Gaza
Deanery of Post Graduate Studies
Faculty of Information Technology

الجامعة الإسلامية – غزة
عمادة الدراســـات العليا
كلية تكنولوجيا المعلومات

Adaptation of Acoustic and Language Model for Improving Arabic Automatic Speech Recognition

A Thesis Submitted to the Faculty of Information Technology
in Partial Fulfillment of the Requirements for the
Degree of Master in Information Technology

By
Oussama Soliman Enshassi

Supervised By
Prof. Alaa El-Halees

January 2016

بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا هاتف داخلي: 1150

الرقم Ref ...../35/غ.س/

التاريخ Date 2016/01/27

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **أسامة سليمان علي إنشاصي** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** **برنامج تكنولوجيا المعلومات** وموضوعها:

مواءمة النموذج الصوتي واللغوي لزيادة فاعلية التعرف التلقائي على المنطوق العربي

Adaptation of Acoustic and Language Model for Improving Arabic Automatic Speech Recognition

وبعد المناقشة التي تمت اليوم الأربعاء 16 ربيع الآخر 1437هـ، الموافق 2016/01/27م الساعة الثالثة مساءً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

أ.د. علاء مصطفى الهليس	**مشرفاً و رئيساً**	.....
د. أشرف محمد العطار	**مناقشاً داخلياً**	.....
د. محمد عبد اللطيف راضي	**مناقشاً خارجياً**	.....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية *تكنولوجيا المعلومات*/ برنامج تكنولوجيا المعلومات.**

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤوف علي المناعمة

P.O. Box 108, Rimal, Gaza, Palestine fax: +970 (8) 286 0800 فاكس Tel: +970 (8) 286 0700 ص.ب. 108 الرمال، غزة، فلسطين هاتف
public@iugaza.edu.ps www.iugaza.edu.ps

Abstract

Automatic Speech Recognition (ASR) is translation of spoken words into text by computer. ASR technology has been widely integrated into many systems. However, Arabic speech recognition applications still suffer from high error rate, which is mainly due to a variation in speech. Variation in speech leads to a mismatch between the Arabic speech and the trained models.

Variation in speech is a major problem in improving the accuracy of Arabic automatic continuous speech recognition applications. Variability may occur at the phonetic, word, or sentence level. In this thesis, the researcher proposes an approach to adapt acoustic model and language model under limited resource for Arabic speakers. A preliminary work on pronunciation model has also been carried out.

Arabic acoustic modeling has been proposed to overcome the variation in speech under limited resource for Arabic speakers. In our case, if there are several Arabic acoustic models available, we can propose a hybrid approach of interpolation and merging of acoustic model for adapting the target acoustic model. The proposed approaches have proven to be very effective to handle the variability existing in the Arabic speech. The Word Error Rate (WER) was measured for both systems. It was found that the baseline system has the WER equals 13.28% which was significantly decreased to 11.04% in the Enhanced system.

Besides, the researcher proposed interpolation approach for adapting the Arabic language model. The results showed that the baseline system has the WER equals 12.4% which significantly declined to 8.4% in the Enhanced system. In addition, the results showed that applying the hybrid of acoustic approach followed by interpolation language approach achieved considerable improvement of 5.32% in the WER. The baseline system has the WER equals 13.28% which was significantly reduced to 7.96% in the Enhanced system.

However, the proposed phonetic rules in pronunciation model did not lead to a significant improvement.

Keywords: Arabic automatic speech recognition, Acoustic modeling, language modeling, Modern Standard Arabic.

الملخص

مؤائمة النموذج الصوتي واللغوي لزيادة فاعلية التعرف التلقائي على المنطوق العربي

إن التعرف التلقائي على الكلام (ASR) عبارة عن ترجمة الكلمات المنطوقة إلى نص من خلال الحاسوب. فقد تم دمج هذه التقنية في العديد من النظم؛ ومع ذلك، فإن تطبيقات التعرف على الكلام باللغة العربية ما زال يشوبها نسبة عالية من الخطأ؛ الأمر الذي يُعزى أساساً إلى الاختلاف في نطق الكلمة الواحدة. فالاختلاف في نطق الكلمة الواحدة يؤدي إلى عدم التوافق بين الخطاب العربي والنماذج المدربة لذلك الغرض.

إن التباين في نطق الكلمة الواحدة لهو مشكلة رئيسه في تطوير دقة تطبيقات التعرف التلقائي على الخطاب المستمر باللغة العربية. فقد يحدث التباين على مستوى اللفظ، أو الكلمة، أو الجملة.

ففي هذه الأطروحة يقترح الباحث أسلوباً لتكييف نموذجاً صوتياً و نموذجاً لغوياً في ظل موارد محدودة للناطقين باللغة العربية. كما تم اجراء عمل مبدئي على القاموس الصوتي.

وتم اقتراح النمذجة الصوتية العربية للتغلب على التباين في نطق الكلمة الواحدة في ظل موارد محدودة للناطقين باللغة العربية، فإذا كان هناك العديد من النماذج الصوتية العربية المتاحة يمكننا اقتراح نهجاً هجيناً يعمل على استيفاء نموذج صوتي او دمجه لتكييف النموذج الصوتي المرجو الوصول إليه. وقد أثبت الأسلوب المقترح أنه فعال جداً للتعامل مع التباين الحاصل في الخطاب العربي. وقد تم قياس معدل الخطأ في الكلمات (WER) لكلا النظامين. وقد تبين أن نظام الخط الأساسي لديه نسبة خطأ في الكلمات تساوي 13.28% التى انخفضت بشكل ملحوظ إلى 11.04% في النظام المحسن.

وبالإضافة إلى ذلك، اقترح الباحث أسلوباً لتكييف النموذج اللغوي، وتوصلت النتائج إلى أن نظام الخط الأساسي لديه نسبة خطأ في الكلمات تساوي 12.4% والتي انخفضت بشكل ملحوظ إلى 8.4% في النظام المحسن. وأظهرت النتائج أيضا أن تطبيق الأسلوب الصوتي الهجين يليه أسلوب استيفاء النموذج اللغوي الذى حقق تطوراً ملحوظاً بنسبة 5.32% في نسبة الخطأ في الكلمات. وأن نظام الخط الأساسي لديه نسبة خطأ في الكلمات تساوي 13.28%، والتي انخفضت بشكل ملحوظ إلى 7.96% في النظام المحسن. ومع ذلك، فإن القواعد الصوتية المقترحة في القاموس الصوتي لم تؤدِّ إلى تطور كبير.

كلمات مفتاحية: التعرف التلقائي على الكلام باللغة العربية، النمذجة الصوتية، والنمذجة اللغة، اللغة العربية الفصحى الحديثة.

Acknowledgments

In the Name of Allah, the Most Gracious, the Most Merciful.

First and foremost, I thank Almighty Allah for giving me the inspiration, patience, determination and strength to complete this work successfully

I would like to express my sincere thanks and appreciation to my supervisor Prof.Alaa El-Halees (Deputy Dean of Information Technology Faculty) for his excellent comments on my research that considerably improved this thesis.

Further, my thanks also go to Dr. Basem Ahmed for his help, discussions, and experimental setup for CMU Sphinx-III during my work on this research. Without their help, my work on ASR system will be much more difficult to accomplish.

Also I would like to thank King Fahd University of Petroleum and Minerals (KFUPM), and Prof. Moustafa Elshafei for providing the broadcast news corpus to work this research.

I am greatly indebted to my mother and my father, who are devotedly have spent their lives to motivate their children to pursue their higher education. I thank also my mother for remembering me in her prayers day and night during my research. In addition, I would like to thank dear my brothers.

Also I would like to thank my wife Hanan for her constant love and encouragement and shouldering an extraordinary responsibility for dealing with two kids (Karam and Abd Al-Rahman) during my work on the master degree. I thank my wife and kids for their patience.

Last but not least, for everyone who has helped and supported me to reach this stage; I would say Thank You.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ASR	Automatic Speech Recognition
NLP	Natural Language Processing
MSA	Modern Standard Arabic
HMM	Hidden Markov Model
GMM	Gaussian Mixture Models
HTK	Hidden Markov Model Toolkit
LPC	Linear Predictive Coding
PLP	Perceptual Linear Prediction
MFC	Mel Frequency Cepstral
MFCC	Mel Frequency Cepstral Coefficient
PCA	Principal Component Analysis
LDA	Linear Discriminative Analysis
ANN	Artificial Neural Network
MAP	Maximum a-posteriori
MLLR	Maximum Likelihood Linear Regression
CI	Context Independent
CD	Context Dependent
WER	Word Error Rate
SLM	CMU-Cambridge Statistical Language Modeling
SCTK	Speech Recognition Scoring Toolkit
CMU	Carnegie Mellon University
BCN	Broadcast News
HQ	The Holy Qur'an
LDC	Linguistic Data Consortium
ELRA	European Language Resource Association

CHAPTER 1

Introduction

In this chapter, the researcher gives an overview of the automatic speech recognition. Subsequently, problem statement and objectives are presented. Then, he looks at the importance of the research and methodology of his work. Finally, scope and limitations of the research are presented.

1.1 Overview

Human speech is the natural form of human communication. The main goal of Automatic Speech Recognition (ASR) is to recognize human speech units such as words and sentences using algorithms implemented in a computer [1-3]. In fact, communicating fluently with machines may eliminate traditional handwriting problems and, therefore increase the productivity of people and machines at the same time. ASR belongs to the field of Natural Language Processing (NLP). Gorin in [4] presented that ASR has a long history of research, which includes contributions from many electrical engineers, phoneticians, linguists, computer and data scientists, mathematicians and other researchers. Over the last few decades the accuracy of ASR systems has significantly improved and the recognition tasks have become larger and more realistic. This has also led to the emergence of various ASR-based commercial products, including automatic speech transcription systems, speech and dialog interfaces, etc.

In general, the ultimate goal of these fields is to improve human-computer interaction by designing machines that are indistinguishable from humans in their ability to hear, recognize, and understand speech as well as vocalizing a spoken language. Over years, ASR technology has been increasingly applied to many types of applications such as dictation software [5-7], voiced-based information retrieval [8-11], automatic processing and documentation of audio data [12-14], speech assistant language learning [15-17], language-to-language translation of a spoken sentence [18, 19], and computer control for people with physical limitations [20]. While there are many ASR related

applications, it is generally agreed upon that higher ASR accuracy results in better service [21-23]. Indeed, speech communication with computers and household appliances is imagined to be the dominant human-machine interface in the near future. However, ASR still faces many challenges before it can be employed by everyone everywhere because the machine capabilities developed are still quite primitive compared to their human counterparts.

A wide range of challenges have to be confronted when processing speech so as to maintain the system's accuracy and robustness. Among these challenges is the limited sources of information given to the recognition engine compared to human knowledge. For example, humans use their knowledge of the speaker, the topic and language system to predict the intended meaning. Second, the variability arising from differences among speakers such as age, gender, level of speech, different dialects, regional accents, and more. Also, the emotional state of a single speaker can also increase speech variability and hence affects the system's accuracy. The recognition task is more challenging with certain languages such as morphologically complex languages which tend to have very rich vocabulary. Moreover, the quality of the recording equipment, room acoustic, channel characteristics and the existence of any kind of background noise in the speech can be obstacles hindering achievement of a high accuracy recognition system.

Variability in Arabic speech leads to ambiguity and misrecognition. This ambiguity in words or phonemes hardly can be classified correctly by recognizer. Although many studies have suggested Knowledge-Base and Data-Driven approaches to reduce effects of variability, these approaches cannot cover most variability speech [24]. Knowledge-based methods are based on linguistic standards. These standards are presented as phonetic rules that can be employed to find possible alternative of pronunciation for word utterances. Data-driven method is based on the training corpus to find possibility of the pronunciation variants.

A statistical-based approach of ASR has become widely used in most of the ASR systems. The statistical model assumptions describe a set of probability distributions, some of which are assumed to adequately approximate the distribution from which a particular data set is sampled. Statistical ASR systems have three types of resources for modeling speech at different stages. These resources are acoustic model, pronunciation

model, and language model. The acoustic model defines the basic units of speech. Unit of speech can be phones, phonemes, syllables, and words. Acoustic model is a statistical representation of the phones as fundamental speech units. The pronunciation dictionary (pronunciation model) maps words into sequences of phonemes. The language model represents the grammar of a language. The language model is a statistically based model using unigram, bigrams, and trigrams of the language for the text to be recognized.

Although the Arab world has an estimated number of 250 million Arabic speakers, there has been little research on Arabic ASR when compared to other languages of similar importance such as Mandarin language [25]. That is why this study concentrates on Arabic ASR. In this study, the researcher investigates Modern Standard Arabic (MSA). MSA is used in writing and in most official speech. MSA is the main medium of communication for public speaking and news broadcasting [26].

1.2 Problem Statement

Variability in Arabic speech leads to ambiguity and misrecognition. This ambiguity in words or phonemes hardly can be classified correctly by recognizer. This lead to less satisfactory accuracy and to high WER.

1.3 Objectives

1.3.1 Main Objective

The main goal of this study is to improve the speech recognition accuracy for Arabic speech recognition systems. The aim is to reduce the WER by using an acoustic model offline adaptation method and language model offline adaptation method where variability in Arabic speech under little speech and text data exists.

1.3.2 Specific Objectives

The specific objectives of the work are:

i. To use two Arabic speech corpora.

ii. To propose a method to enhance Arabic ASR to tackle variability of speech and lack of corpus training data based on acoustic model and language model.

iii. To propose new rules for creating pronunciation dictionary.

iv. To train the proposed method based on pronunciation model, language model, and acoustic model.

v. To evaluate the proposed method on a benchmark dataset. The most commonly used metric in evaluating the accuracy systems is the word error rate.

vi. To compare the results obtained from hybrid model performed on broadcast dataset with previous work models used this dataset in order to be sure that my model has achieved its main objective.

1.4 Importance of the Research

The Arabic ASR research has been growing very slowly in comparison to English and other languages ASR research. Conversely, there is a great interest in changing this situation, both from a scientific and economic point of view, given the rising economy in some of the Arab countries.

ASR is crucial for many applications in daily life adding communication easiness in situations where text information is preferred over voice. ASR is a necessary step towards automatic translation between different languages, being essential for worldwide exchange. Additionally, Broadcast news transcriptions enable handicapped people such as deaf to access information from spoken news, providing assistance in classroom situations and facilitating the use of the telephone for hard of hearing people. Furthermore, speech recognition may improve communication of people with language disorders. Last but not least, ASR adds a new dimension to information retrieval by archiving speech as text and making it searchable.

However, big challenges for current ASR systems are regional differences, dialects, age, gender and accents of the speech to be recognized. This thesis addresses this issue with respect to the Arabic language.

1.5 Methodology

The methodology presents the proposed method for improving the accuracy of Arabic ASR systems. The main idea of new approaches is based mainly on offline adapted acoustic model from another acoustic model. Moreover, this study attempts to propose a method based mainly on adapting the language model by another text corpus. The research method consists of five main phases as follows:

Phase1: Acquiring of an Arabic speech corpus.

This phase will use Arabic broadcast news speech corpus in order to build Arabic ASR system. Then, we can train and test ASR models.

Phase2: Data preparation.

This phase will generate a control file, transcription files, filler file, and phone file.

Phase3: Building a pronunciation model.

In this phase, the pronunciation rules will be designed in order to generate phonetic transcription for all vocabulary in this study using Java program.

Phase4: Building language model.

The language models generated from the diacritized transcription of Arabic broadcast news speech. It will be built using CMU-Cambridge statistical language modeling toolkit.

Phase5: Building an acoustic model.

This phase will use SphinxTrain software to create acoustic model. The Sphinx trainer is based on Hidden Markov Model.

Phase6: Evaluating baseline of Arabic ASR system.

We will record word error rate as baseline for our work. This WER will compare with other results to check whether the accuracy is improved.

Phase7: Building another speech corpus.

The aim of this phase is to create another acoustic model (source acoustic model) from another speech corpus. Then, we will use the source acoustic model for testing the proposed methods in adaptation process.

Phase8: Adaptation of an acoustic model and a language model.

The general approach of combining two acoustic models is proposed to model variations in speech which exist among Arabic speakers in a simple and robust manner.

The aim of combining the two acoustic models method is to benefit from phones that have been training more accurate in another acoustic model. Each phoneme is represented by a state, and each state is defined by value with a single Gaussian.

The second adaptation is language model interpolation. The aim is to add new sources of information to the previously existent models with the objective of enriching them. The goal in language model adaptation is to reflect the changes that the language experiences when moving towards different domains. The language model interpolation consists of taking a weighted sum of the probabilities given by the component models.

Phase9: Decoding the data (decoder).

We used Spinx3 decoder engine with testing data to generate transcription file. The decoder will be carried out based on three models which are acoustic, language, and pronunciation model.

Phase10: Evaluation.

WER is a common evaluation metric of the accuracy of ASR systems and many other natural language processing tasks. The WER is based on the transcription file that was generated from the decoder.

Phase11: Comparing.

Comparing the WER obtained from adapted model which performed on broadcast dataset with previous related works models used this dataset in order to be sure that our model has achieved its main objective.

1.6 Scope and Limitations

The scope of the research to be conducted in the project research covers acoustic and language offline adaptation technique for continuous speech recognition, with independent speaker as the test domain. Also, ASR system is based on statistical model. In addition, the recognition focuses on broadcast news, and this study focuses on continuous MSA speech data with transcript and diacritical marks. Finally, the proposed method considers the words related to news domain from an Arabic corpus.

Conversely, my research does not focus on different dialects and regional accents, and emotional state in the project research. In addition, the research project does not pay attention for informal Arabic speech and writing in our experiments. In addition to the

above limitation, there is background noise, interference from other speakers, room acoustics, recording equipment, and channel characteristics in the case of telephone conversation that is not considered in the project research. Finally, we do not consider spontaneous speech and dependent speaker in the study.

1.7 Contributions

The main contribution of this study is the improvements achieved in Arabic ASR over the baseline system. These improvements are carried out by utilizing hybrid of acoustic model interpolation and merging approach, and interpolation language model. Our results present the following findings:

- Hybrid acoustic model interpolation and merging approach, which is based on creating intermediate acoustic model between target and source acoustic models, leads to important improvements in the WER.
- Improved hybrid acoustic model interpolation and merging approach leads to significant improvements in the WER.
- Interpolation language model approach leads to significant improvements in the WER.
- Cross-validation test is employed for training and testing purpose. This new method can be used to determine a reasonable value for the accuracy.

1.8 Thesis Outline

This thesis is structured as follows: Chapter two is the literature review and the background of this research work. Chapter three shows the related work. Then, in chapter four, the proposed method is described. Chapter five includes experiments and results. Finally chapter six presents the conclusions and future works.

CHAPTER 2

Literature Review

In this chapter, we first briefly present the general background of an ASR system. Subsequently, we look at the architecture of statistical automatic speech recognition system and its components. Finally, recognition evaluation is presented.

2.1 Background

ASR systems are software that convert speech into texts for many purposes such as the ability of machine to understand speech. Nowadays, humans can verbally communicate with computer. Indeed, handwriting problems may be eliminated by speaking fluently with computer, and, therefore, increases the productivity of the people. However, before voice can be analyzed for its meaning, it has to be first converted to a simpler form (the text transcription). Speech recognition or speech to text is an interesting but challenging domain because of its multidisciplinary research area. Among the domains involved are signal processing, pattern recognition, linguistics, information theory [27].

A typical large vocabulary speech recognizer would first convert speech waveform into a sequence of feature vectors to be used to identify the phones (the acoustic speech unit). The recognized phones are used to specify the words and then the sequence of words.

The statistical method has dominated ASR research over the past few decades [28]. The statistical method is dominated by the powerful statistical based approach which uses Hidden Markov Model (HMM). Several researchers in [29-32] have stated that the HMM-based ASR technique achieved various successful applications requiring large vocabulary speaker-independent continuous speech recognition.

The HMM-based method contains recognizing speech by estimating the likelihood of each phone at contiguous, small frames of the speech signal [33, 34]. Words in the target vocabulary are converted to a sequence of phonemes and then a search process in

the words in the vocabulary list is used to find the phoneme sequence that best matches the sequence of phones of the spoken word. All phonemes are modeled as a sequence of HMM states. Standard HMM-based systems provide the likelihoods of a certain frame observation being produced by a state. The likelihoods are estimated using traditional Gaussian mixture models. The likelihoods are also known as the emission probabilities. There are several advantages of using HMM with Gaussian Mixture Models (GMM) such as efficient learning and decoding algorithms, a rich mathematical framework, and an easy integration of multiple knowledge sources [27].

The HMM tools, also known as the Hidden Markov Model Toolkit (HTK) [35, 36], and the CMU Sphinx system [37, 38] are common tools that are used in developing speaker independent, large vocabulary, and continuous speech recognition. HTK toolkit is used to build HMMs. On the other side, Sphinx system is built only for speech recognition systems. This study is based on Sphinx-based ASR system for testing and evaluation.

Several researchers in [39-41] have presented that continuous HMM, and semi-continuous HMM technique can be used in CMU Sphinx 3 for parametrizing the probability distributions of the state emission probabilities. The continuous HMM technique is slower in decoding and requires more parameters. However, the semi-continuous technique is faster in decoding and uses substantially a smaller number of parameters. Although the continuous HMM technique shows to be effective for large vocabulary applications, the semi-continuous technique is only good for limited vocabulary.

2.2 Classification of ASR System

The speech recognition systems can be categorized into several different classes such as isolated words, connected words, continuous speech and spontaneous speech [42].

First, isolated words (tens of words) recognizers usually require each utterance to have quiet. It accepts single words or single utterance at a time. These systems require the speaker to wait between utterances. Second, connected words systems are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause

between them. Third, continuous speech (thousands of words) recognizers allow users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries. Fourth, spontaneous speech (tens of thousands of words) can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together like "ums" and "ahs", and even slight stutters [43].

The isolated word recognition was carried out in the ASR area resulting in numerous practical and commercial successes with notable high recognition accuracy. In contrast, continuous speech recognition has provided many issues to ASR systems [44].

In this thesis, the researcher proposed methods to improving recognition accuracy of continuous speech for Arabic language.

2.3 An Automatic Speech Recognition System Architecture

An ASR system is also called speech to text system. The purpose of ASR systems is to convert an utterance as input into an output text transcription. ASR system consists of decoding and training modules [45]. **Figure 2-1** illustrates the main components of a large vocabulary continuous speech recognition system and the information flow through it [46].

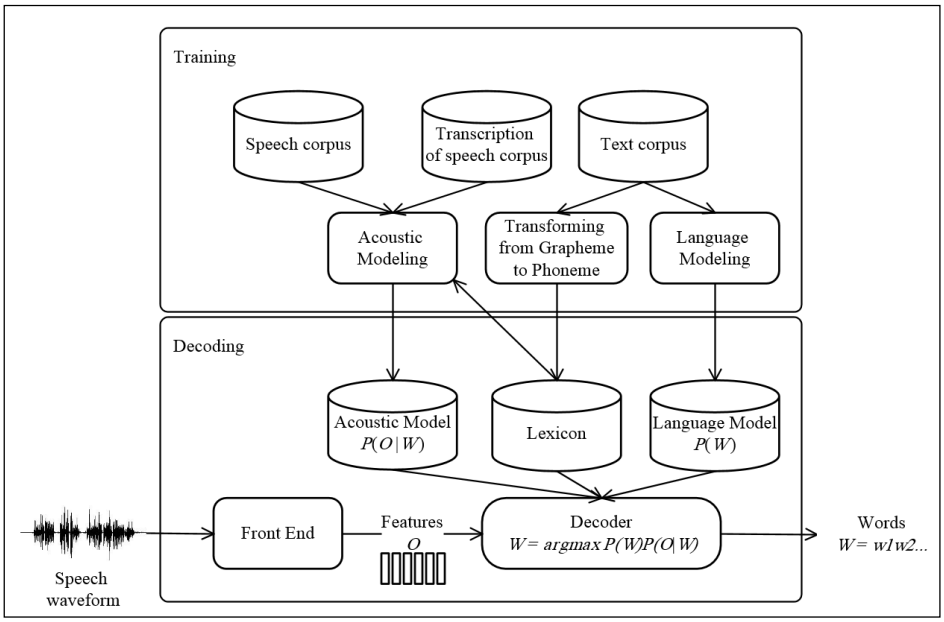

Figure 2-1 Components of ASR system and the information flow through it.

In general, the decoding components consist of a signal processing front-end and a decoder. The purpose of signal processing front-end is to digitize analog signal and to convert it to discriminative features for recognition. The decoder is the engine of a speech recognition system that reveals the possible word sequence from the feature vectors using the knowledge from pronunciation model, acoustic model, and language model. From the point of view of grammatical, these models are generally expressed in language as follows: acoustic model seems like phonology of a language, pronunciation model seems like vocabulary and pronunciations, and language model seems like grammar of a language [46].

Consider an Arabic sentence "أَهلاً بِكُم مِن جَدِيد" which means in English "Welcome back", recorded in a wave file. It shows that such a simple speech message contains various levels of information. These levels and the corresponding models are shown in **Figure 2-2** based on [4].

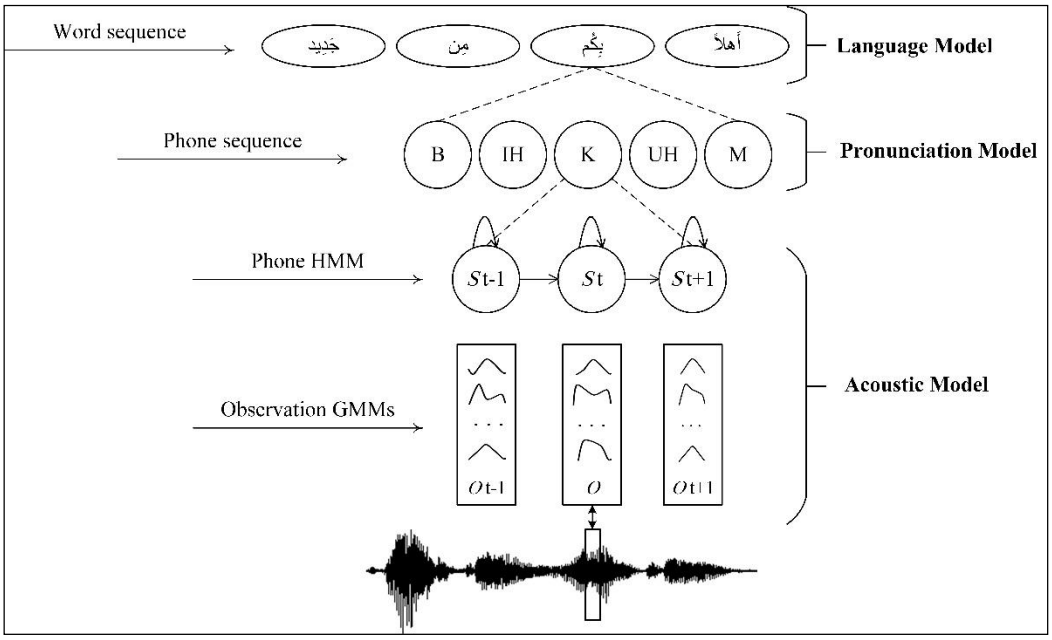

Figure 2-2 Modeling different levels of information of a spoken sentence by a HMM

The acoustic model defines the basic units of speech. Unit of speech can be phones, phonemes, syllables, and words. Acoustic model is a statistical representation of the phone. However, pronunciation model (or lexicon language or dictionary) represents mapping each word into sequences of phonemes. Lexicon also provides units such as word or syllable. The language model defines the syntax and structure of a language with the vocabulary from the pronunciation dictionary. In detail, the ASR system consists of the following:

2.3.1 Front-End Signal Processing

The signal processing front-end stage aims to extract discriminative features that are perceptually important. The front-end signal processing first digitizes the analog signal to appropriate pattern for analysis. The processing contains several phases such as pre-emphasis, filtering, sampling, and quantization [47]. A digitized single speech at a sampling rate of 16 kHz is sufficient to represent human speech intelligibly. Higher sampling frequency does not contribute any more improvement to the speech recognition system [48].

Next, the digitized signal or speech signal is converted to feature vectors. The possible kinds of feature are short time spectral envelope, energy, zero crossing rates, and level crossing rates. Frequency-domain features such as short time spectral envelope are more accurate and descriptive compared to time-domain features for analyzing speech. The spectral analysis provides several methods such as: Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), and Mel-Frequency Cepstral (MFC) spectral analysis model. One of the most widely used features in speech recognition is the Mel Frequency Cepstral Coefficients (MFCC) [49, 50].

Researches show that sufficient data to represent speech is 13th-order MFCC, for example 39 coefficients [48]. In addition, to the raw MFCC features, the first and second derivatives of the MFCC features are normally computed, because they provide temporal changes of information of the spectral. For speech recognition system using HMM, this information can be useful, because the acoustic frames are assumed to be independent and stationary [48]. Principal Component Analysis (PCA) and Linear Discriminative Analysis (LDA) techniques are used to reduce the size of the feature vectors by applying them on the vectors to create a more compact and discriminative feature [48].

2.3.2 Decoder

The word decoder is developed from the information theory domain. The word decoder means the conversion of a coded message to an understandable form. The decoder in speech recognition uses the speech features presented by the Front-End to reveal the most probable word sequences and, then, sentences that correspond to the speech signal or more precisely the feature vectors. The search for the most probable word sequence can be achieved by maximizing the posterior probability for the given feature vectors. It is difficult to calculate efficiently and robustly the posterior probability. Thus, instead of calculating the posterior probability directly, it can be put in another form using Bayes rule as seen in equation 1 from [4, 48]:

$$\widehat{W} = \underbrace{\arg max}_{word \in L} P(W|O)$$

$$\widehat{W} = \underbrace{\arg max}_{word \in L} \frac{P(O|W)\,P(W)}{P(O)} \quad (1)$$

$$\widehat{W} = \underbrace{\arg max}_{word \in L} \underbrace{P(O \mid W)}_{Acoustic\ Model} \underbrace{P(W)}_{Lnaguage\ Model}$$

The **Equation (1)** is the most direct way to calculate all the possible word sequences and select the one which gives the highest value from the formula. Where $\widehat{W}$ is most probable word sequence of the spoken words w_1, w_2, …w_m which gives the maximum posterior probability P(W|O) given O, a series of observations o_1, o_2,… o_n which produce the word sequence, and L denotes a language. So, the best word sequence can be found by combining the language probability of word sequence $P(W)$, which is called prior probability, with the acoustic probability of the word sequence $P(O|W)$, which is also called conditional probability, from an acoustic model which gives the highest value. In automatic speech recognition, the state of the art acoustic model used hidden Markov model [48]. On the other hand, the language model provides the language probability. A widely used language model is the n-gram model [48].

2.3.3 The Acoustic Modeling

After years of research and development, accuracy of automatic speech recognition remains one of the most important research challenges. A number of well-known factors determine accuracy; those most noticeable are variations in context, in speaker, and in environment. Acoustic modeling plays a critical role in improving accuracy and is arguably the central part of any speech recognition system [48].

The acoustic model defines the basic units of speech. Unit of speech can be phones, phonemes, syllables, or words. The Acoustic model is a statistical representation of the phones as fundamental speech units. One of the main challenges in the field of speech recognition is building a robust acoustic model. The variability which exists in the speech is the main difficulty of modeling acoustic features in a robust manner. The context variability may occur at the phonetic, word or sentence level [48]. In continuous speech, words in a sentence may be connected rather than separated by a silent period.

Furthermore, variability may occur in pronunciation when the words are pronounced in a separate and continuous style. At the phonetic level, variability may also occur in a phoneme when it is realized under different contexts. Variation in speaker level is the most significant since the speech is influenced by the physical properties such as age, gender, vocal tract size, and also social characteristics. Another contribution to speech variability is environmental condition. Therefore, the accuracy of a speech recognition system can influence the environmental noise and variation in microphone.

There are many possible approaches for modeling the acoustic units such as: HMM as in [51-55], Artificial Neural Network (ANN) as in [56-59] and template model [60]. HMM approaches become widely popular in the last several years in statistical speech recognition because of its robustness [46]. In this study, the researcher selected the HMM approach.

The HMM theory has been developed since the late 1960s. HMM has been employed in automatic speech recognition by IBM since 1970s. A Markov chain is a stochastic procedure with short memory, where the current state depends on the previous state only. In a Markov chain, the observations are actually the state sequence. The HMM is an extension of a Markov chain where the observation is a function of the state; therefore, the state sequence is hidden in HMM. The probability to be at a particular state can be calculated instead given the observation [48].

The acoustic model is a statistical representation of the phone. One of the key factors to improve recognition accuracy is acoustic model as it characterizes the HMM of each phone. The CMU pronunciation dictionary in [61] presented that CMU Sphinx contains 39 English phonemes. The acoustic model can use a 3 to 5 state Markov chain in order to represent the speech phone [37]. **Figure 2-3** illustrates a 3-state phone's acoustic model [62]. S1 denotes the representation of phone at the beginning, while S2 and S3 denotes the representation of the phone at the middle and the end states, respectively. S1, S2, and S3 are mixture Gaussian densities that describe the behavior of the feature vectors of the phone.

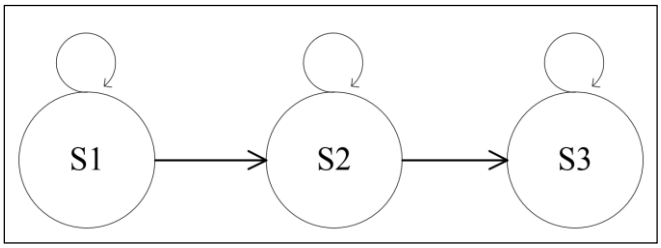

Figure 2-3 A Simple 3-state phone left-to-right HMM topologies

Hidden Markov model is defined by the following set of parameters [33]:

- *N* denotes the number of states.
- A denotes the state transition probabilities, $a_{ij} = P(s_{t+1} = j | s_t = i)$, where s_t is the state at time *t*.
- B denotes the observation symbol probability, $b_j(x_t) = P(x_t | s_t = j)$, where x_t is the observation at time *t*.
- Π denotes the initial state probabilities. $\pi_i = P(s_1 = i)$

Each phoneme in continuous speech is influenced in several levels by its neighboring phonemes. Therefore, Sphinx[1] utilizes triphones for better acoustic modeling. Triphones are context dependent models of a sequence of three phonemes. A phoneme surrounded by exact left and right phonemes is represented by each triphone. For example, the phoneme /n/ when /ae/ appears on its left and /d/ appears on its right is the triphone /n(ae, d)/ (END) [63].

Continuous HMM uses the Gaussian mixture density. The probability of producing the observation x_t given the transition state *j*, $P(x_t | j)$ is given in equation 2 [48].

$$b_j(x_t) = p(x_t | q_t = j) = \sum_{k=1}^{M} w_{j,k} N_{j,k}(x_t) \qquad (2)$$

[1] Sphinx is the general term to describe a group of speech recognition systems developed at Carnegie Mellon University. These include a series of speech recognizers (Sphinx 2 - 4) and an acoustic model trainer (SphinxTrain).

where $N_{j,k}$ is the k-th Gaussian distribution, $w_{j,k}$ are the mixture weights, and $\sum_k w_{j,k} = 1$. Continuous HMM is the most common technique for large vocabulary speech recognition systems. Conversely, its main disadvantage is that a very large number of parameters required to describe the Gaussian distributions.

An efficient number of parameters to describe all triphones in acoustic model can be achieved by employing the concept of shared distributions. The shared distributions method aims to provide a common pool of probability distributions for all triphones of a given phoneme which are called Senones [63].

Figure 2-4 show that three phases are included in the training procedure for the acoustic model [64]. Each phase occurs in three steps: model definition, model initialization, and model training. Each phase makes use of the output of its previous step. The following phases are:

In the first phase, Context-Independent (CI) phase generates one HMM for each phoneme in the phoneme list. The number of states in an HMM model can be specified by the developer; in the model definition stage, a serial number is assigned for each state in the whole acoustic model. Additionally, the main topology for the HMMs is created. The topology of an HMM specifies the possible state transitions in the acoustic model, the default is to allow each state to loop back and move to the next state; however, it is possible to allow states to skip to the second next state directly. In the model initialization, some model parameters are initialized to some calculated values. The model training stage consists of a number of executions (5 to 8 times) followed by a normalization process.

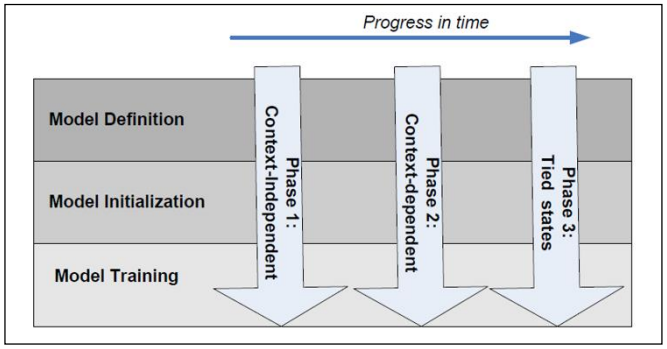

Figure 2-4 Building the acoustic model phases

Untied Context-Dependent (CD) phase: In the second phase, triphones are added to the HMM set. In the model definition stage, all the triphones appearing in the training set will be created, and then the triphones below a certain frequency are excluded. Specifying a reasonable threshold for frequency is important for the accuracy of the model.

After defining the needed triphones, states are given serial numbers as well (continuing the same count). The initialization stage copies the parameters from the CI phase. Similar to the previous phase, the model training stage consists of number of executions of Baum-welch algorithm followed by a normalization process.

In the third phase, tied context-dependent phase aims to improve the accuracy of the model generated by the previous phase by tying some states of the HMMs. These tied states are called Senones. The process of creating these Senones involves building some decision trees.

In this research work, the researcher used the Sphinx 3 default setting. After the new model is defined, the training procedure continues with the initializing and training stages. The training stage for this phase may include modeling with a mixture of normal distributions. This may require more iterations of Baum-welch algorithm.

2.3.4 The Language Modeling

Language model also can represent pronunciation variants (variation speech). The language model represents the grammar of a language. ASR systems treat the recognition process as one of Maximum A-Posteriori (MAP) estimation, where the most likely sequence of words is estimated, given the sequence of feature vectors for the speech signal. A mathematical estimation is defined by [65]:

$$\begin{aligned} &Word_1\, Word_2 \ldots Word_n \\ &= \arg\max_{w1\, w2 \ldots}\{P(feature\ vectors|w_1\, w_2 \ldots)P(w_1\, w_2 \ldots)\} \end{aligned} \tag{3}$$

where $Word_1 Word_2 \ldots Word_n$ is the recognized sequence of words and $w_1 w_2 \ldots$ is any sequence of words. The argument on the right hand side of **Equation (3)** has two components: the probability of the feature vectors, given a sequence of words

$P(feature\ vectors|w_1\ w_2\ \dots)$, and the probability of the sequence of words itself, $P(w_1\ w_2\ \dots)$. The first component is provided by the acoustic model. The second component, also called the language component, is provided by a language model. The most commonly used language models are n-gram language models. These models assume that the probability of any word in a sequence of words depends only on the previous n words in the sequence. Thus, a bigram language model would compute P(W1 W2 ...) as [48]:

$$P(W_1\ W_2\ W_3\ \dots) = P(W_1)P(W_2|W_1)P(W_3|W_2)\ \dots \quad (4)$$

Similarly, a trigram model would compute it as

$$P(W_1\ W_2\ W_3\ W_4\dots) = P(W_1)P(W_2|W_1)P(W_3|W_2W_1)P(W_4|W_3W_2)\ \dots \quad (5)$$

The available transcription text can be used to build the model is called the training corpus [48]. The n-gram language model is trained by counting n-gram occurrences in a large transcription corpus to be then smoothed and normalized. In general, an n-gram language model is constructed by calculating the following probability for all combinations that exist in the transcription corpus:

$$P(w_1^n) = \prod_{k=1}^{n} p\left(w_k|w_1^{k-1}\right) \quad (6)$$

where n is limited to include the words' history as bigram (two consequent words), trigram (three consequent words), 4-gram (four consequent words), etc. for example, by assigning n=2, the bigram is calculated for the words sequence as $(W_1\ W_2\ ) = p(W_1)p(W_2|W_1)$.

Clarkson and Rosenfeld in [66] described the CMU statistical language tool. The CMU statistical language toolkit is used to generate our Arabic statistical language model. The steps for generating and testing the language model is shown in **Figure 2-5** [66]. First, it computes the word unigram counts. Next, it converts the word unigram

counts into a vocabulary list. Finally, it generates bigram and trigram tables based on this vocabulary.

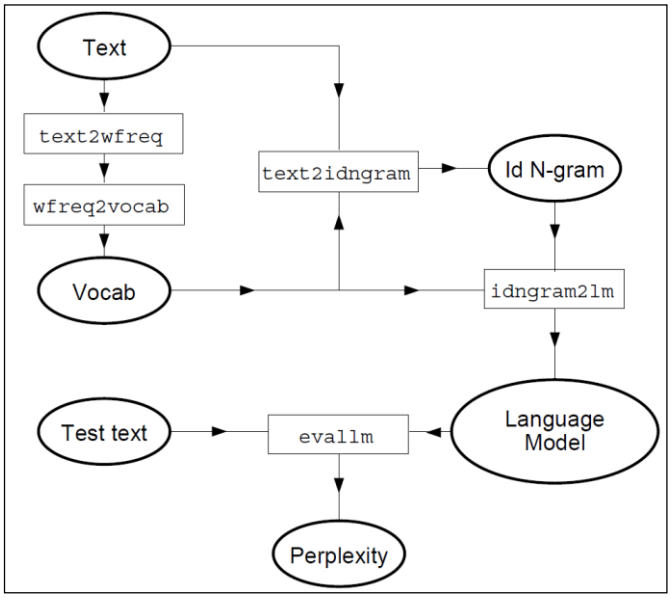

Figure 2-5 Steps for generating and testing the language model

The tool generates the language model in two formats; a binary format to be used by the Sphinx decoder, and a portable text file in the standard ARPA format. In addition, the language modeling tool provides a tool for evaluating the language model which is named perplexity. In this study, we use this tool for building ASR language model.

2.3.5 The Pronunciation Modeling

A pronunciation dictionary (lexicon) is required in training and recognition phases. A pronunciation dictionary maps words into sequences of phonemes. A particular set of words are used within a pronunciation dictionary. A pronunciation dictionary contains the pronunciation of all vocabulary in the text corpus using the defined phoneme set. The accuracy of the ASR systems depends on the dictionary and its phoneme set like acoustic model and language model. In decoding phase, the dictionary assists as an intermediary between the acoustic model and the language model. Closed vocabulary and open vocabulary are two types of dictionary. In closed vocabulary, all corpus transcription words are listed in the dictionary. However, it is possible to have non-corpus transcription words in the open vocabulary dictionary. Typically, a phoneme set,

that is used to represent dictionary words, is manually designed by language experts [67]. However, Clarkson and Rosenfeld in [66] demonstrated that when human expertise is not available, the phoneme set is possible to be selected using a data-driven approach. In addition to providing the words phonemic transcriptions of the target vocabulary, the dictionary is the place where alternative pronunciation variants are added.

2.4 Recognition Evaluation

Accuracy of an automatic speech recognition system is usually estimated by primary metric: Word Error Rate (WER). The task of evaluating a speech recognition system involves comparing a reference (or correct) word sequence with the hypothesis word sequence returned by the ASR system. In such systems, the goal is to decrease the WER.

The WER is estimated from three types of errors in a speech recognition system which are insertion, substitution, and deletion. Insertion refers to addition of an extra word. Substitution refers to replacement of a correct word by another different word. Deletion refers to elimination of a correct word.

For estimating the number of insertions, substitutions, and deletions made, the hypothesis from the system is aligned against the actual reference transcription using a minimum number of single-character edits (Levenstein distance), with the same cost given to the errors ,which can be calculated using equation (7) [27].

$$WER = \frac{N_{Inseration} + N_{Substitutions} + N_{Deletions}}{N_{words\ in\ the\ correct\ sentence}} * 100\% \quad (7)$$

where:

$N_{Inseration}$ denotes the number of the insertions words errors,

$N_{Substitutions}$ denotes the number of substitutions words errors,

$N_{Deletions}$ denotes the number of the deletions words errors,

$N_{words\ in\ the\ correct\ sentence}$ denotes the number of words in the testing set.

Also, the word accuracy can be measured using WER as the following formula:

$$Word\ Accuracy = 1 - WER \quad (8)$$

CHAPTER 3

Related Work

This chapter is dedicated to a brief overview of state-of-the-art of ASR techniques. The first subsection presents common strategies to deal with the adaptation of acoustic models. Subsequently, the researcher looks at common strategies to deal with the adaptation of language and pronunciation models.

3.1 Acoustic Model Adaptation Approaches

It is very hard for humans to reproduce the same exact action twice. Therefore, speakers cannot articulate the speech sound in the same way, because their speech is often influenced by their accent. Speakers from different origins often have pronunciation habits related to their accent language. One of the most important issues is variability in speech which is known to reduce recognition accuracy. Mehla et al. in [42] demonstrated that each person has a different vocal tract, controlled by a unique brain, therefore a large range of variability exists in speech signals. In addition, humans cannot reproduce the same exact action twice; even when attempting to repeat a word uniformly, slight variations occur. Seigel et al. in [68] presented that the complexity of a speech recognition task is expressed through the definition of a number of important issues such as: the vocabulary size, the speaking style, speaker variability, the acoustic environment and channel conditions.

Obtaining adequate speech to generate suitable acoustic model can be considered as time consuming and sometimes unfeasible. Furthermore, the coverage of any corpora cannot contain complete information about all aspects of language lexicon and grammar [3], due to the limited written training data and therefore inadequate spoken training data. One of the main problems in Arabic ASR research is lack of spoken and written training dataset [51]. Al-Sulaiti and Atwell in [69] presented that the most common list corpora between 1986 and 2005 provided only 19 corpora which are 14 written, 2 spoken, 1 written and spoken, and 2 conversational. These corpora are not easy to

provide to the public and many of them can only be obtained by purchasing from the European Language Resource Association (ELRA) or the Linguistic Data Consortium (LDC). Clearly, there is a shortage of spoken data as compared to written data resulting in a great need for more speech corpora in order to serve Arabic ASR. The available spoken dataset was mainly collected from news broadcasting, and telephone conversations.

The significant factor to improve recognition accuracy is the accurate acoustic model as it characterizes the HMM of each phone [27]. Therefore, the current approaches proposed to adapt the acoustic model to enhance the ASR. The Acoustic model has been adapted in ASR in different ways, for example reconstruction model, interpolation model, merging, and hybrid models.

3.1.1 Acoustic Model Reconstruction

The acoustic model training is the most straightforward way of generating an acoustic model. However, it is not easy to acquire adequate MSA speech to generate MSA acoustic model. Therefore, rather creating it from scratch, existing target language acoustic model is employed as the starting point model, which will be later adapted using some acquired MSA speech. Several studies have shown that native speech can be useful for adapting the target language of non-native speakers acoustic model where non-native speech is not available by several authors [70, 71].

Several studies in [46, 72] non-native speech have presented that there are three kinds of resources that can be assisted to adapt the target acoustic model. First, L1 is the native language of the speaker. Second, L2 is any non-native language spoken by the same native group. Third, L3 is language close to the native language of the non-native speaker. For example, if we consider French language as the target language for ASR system, and if the task is to recognize non-native speech from Vietnamese speakers, the resources considered will be Vietnamese speech (L1), any non-native speech uttered by Vietnamese for example non-native English by Vietnamese (L2) and a language close to Vietnamese (L3), respectively [46]. Consequently, this study uses MSA (L1) to adapt the target MSA acoustic model.

Other approaches using phonetic decision trees to generate tied-states throughout training have been suggested to enhance the accuracy of accented and ASR systems for non-native speakers. The idea is to initialize acoustic features at state tying. These methodologies have decreased the recognition errors of non-native speech and at the same time cause little or no decrease in the accuracy of native speakers. This means that the same model can be applied for both native speakers and non-native of Arabic speakers at the same time. Oh et al. in [73] demonstrated an adaptation in the training process of acoustic models. This is done by utilizing a standard spoken language ASR system to create an intra-language "phonetic confusion" matrix between spoken language phones by using decision tree. This matrix is then employed to tie the triphone models of the confused phones during their training.

In addition, a slightly different model has been proposed for tying states for accented speech by using decision tree by several authors [74, 75]. There are two types of tree that are being used. Firstly, a standard phonetic decision trees built using the target language. Secondly, auxiliary tree which is also a phonetic decision trees but built with non-native data, where a particular phoneme is pronounced as another phoneme. The leaves of the auxiliary trees with single Gaussian density each will be merged to the nearest leaf nodes of the standard target language phonetic decision tree by applying weights. The goal is to initialize the densities that define non-native speech. Consequently, standard training processes are as follows, where only native speech is used for creating the acoustic model.

3.1.2 The Acoustic Model Interpolation

Reconstruction of acoustic model needs the raw speech data for modeling, whereas interpolation of acoustic model can be achieved even when the sources are in the acoustic models form. The acoustic model interpolation is usually achieved by applying weights between two acoustic models. Witt in [76, 77] has presented that a target language acoustic model may be interpolated with the native language acoustic model of the speaker. The target language of acoustic model achieved good results for experiments in which the speakers' source language was Japanese or Spanish and the target language was English. He uses algorithms based on hypothesis that there are

intermediate languages between native and non-native speakers. Three approaches for finding the target and source language model mapping have been proposed. Firstly, the using of linguistic knowledge for mapping the target and source language sounds. Secondly, possibility is to conduct perception analysis by phonetician. Finally, using some non-native speech to create confusion matrix in order to find the phoneme confusion.

Alternatively, interpolating of two acoustic models can be between target language and the native language acoustic model. Several researchers have suggested to interpolate the target language acoustic model with the non-native acoustic model [71, 78]. They are created models with only limited amount of non-native speech. However, the interpolation between target and source of the Arabic acoustic model has never been done before.

Moreover, Steidl et al. in [79] considered that acoustic models of native speech are sufficient to adapt the speech recognizer to the way how non-native speakers pronounce the sounds of the target language. In a continuous HMM acoustic model, the matching Gaussians in the corresponding states will be interpolated. If the source acoustic model for interpolation is not derived from the target language acoustic model, the Gaussians in the target and source are mismatched, so they have to be matched first before interpolation can be carried out using distance measure. The interpolation that achieved the best score is actually performed.

Tan and Besacier in [80] demonstrated three interpolation methods based on the use of both the target language acoustic models and the mother tongue acoustic models of nonnative speakers. These three models are manual interpolation, weighted least square based interpolation, and eigenvoice based interpolation. The three acoustic model interpolation methods consist of two identical steps for preprocessing and one different step for the acoustic model interpolation. However, the researcher will use manual interpolation between two of Arabic acoustic models.

Moreover, to improve the accuracy of an ASR system on non-native speech, an interpolation technique between two acoustic models (that of the native language of the speaker or "source language" and that of the spoken language or "target language") has

been proposed in [81]. However, the proposed solution is limited to cases where the spoken language of the speaker is Arabic.

The benefits of acoustic model interpolation are that the method can be simply achieved and the interpolated model has the same number of Gaussians and states.

3.1.3 The Acoustic Model Merging

Similarly, acoustic model merging needs only the acoustic models, without any raw corpus. Merging of acoustic models includes combining two or more acoustic models from normally two resources. Several studies in [76, 77, 82-84] presented that a new model is constructed from combining the target language acoustic model with the corresponding native language acoustic model of the non-native speaker. The idea is that different speakers are possible to use different strategies to pronounce a sound. In this case, it may be the target language speech sound or the speaker's native language sound. Minematsu et al. in [85] demonstrated merging two acoustic models which are the native and the non-native models. A weight will be assigned to each of the merged model, either on the transition of each model to form a new model with six states or into the mixture weights to form a new model with 3 states. Bouselmi et al. in [84] stated that weights can be applied manually or estimated automatically using some speech. However, there are drawbacks associated with the acoustic merging. One of them is that it increases the number of states or Gaussians in each HMM. Second is that it can create some redundant distributions, which therefore increase the memory and computation time.

3.1.4 Hybrid Approach: Acoustic Model Interpolation and Merging

Tan and Besacier in [81] demonstrated a hybrid approach of interpolation and merging of acoustic models of the native languages of the speakers (L1) with those of spoken languages (L2). They presented that there are three possible resources for adapting the target language acoustic model. They are namely the native language (formal Language) of the speaker (L1), any non-native language spoken by the same native group (L2), and languages close to the native language of the non-native speaker (L3) [46]. Their approach does not require extra non-native speech in the process of

adapting two acoustic models. And their approach is based on the acoustic models of L1 and L2. However, they did not consider adaptation between two native acoustic models and Arabic language in their experiments. In addition, Basem and Tan in [86] proposed also a hybrid approach of interpolation and merging to adapt acoustic models of different languages to recognize speech which contains more than one language. The adapted acoustic models showed reduction in WER.

For the mentioned studies, the proposed method is used to adapt two different native and non-native speaker acoustic models in different ways in order to enhance the ASR for native or non-native speakers. However, none of these studies considered modeling the Arabic language as the source or target language. Within the present context and to the best of the researcher's knowledge, this work represents the first using of the hybrid acoustic model approach for improving native Arabic speech using another native Arabic speech. Therefore, this study proposes to adapt the acoustic model for Arabic language. In this study, the L1 is examined for Arabic speaker's adaptation.

3.2 Language Model Adaptation Approaches

Acquiring enough MSA speech to generate MSA acoustic model is a more difficult mission. Also obtaining sufficient MSA speech to model the grammars of Arabic speakers is not an easy task. Over the past years, the amount and diversity of information available online has exponentially grown and this tendency appears to remain unaltered in the near future. As a result, the quality of language models has increased in certain domains where such data became available. Nevertheless, this behavior seems to be reaching an upper limit and it is possible that this continuous increase of information does not lead to any significant improvement in language models [87]. For this reason it is important to find new sources of information that increase the capacity of the data to describe and model the type of language that is being used in an automatic speech recognition application.

For an optimal adaptation of language models in specific domains, it is required that the system has a previous knowledge of data belonging to the same domain or, at least, to a related one. Indeed, the main goal in statistical language model adaptation is to add new sources of information to the previously existent models with the objective of

enriching them. Bellegarda in [88] presented that the aim in language model adaptation is to reflect the changes that the language experiences when moving towards different domains or, as in some applications, when dealing with multiple speakers.

As long as the sources of information for generating language models remain the same, the model will remain static. That is, regardless of the addressed topic, domain or style, the probability of events will not change. However, a static model is not the best option for modeling language in multi-topic speech. In a natural conversation between humans, the topic, subject, genre, style change often, and therefore the word usage changes accordingly. For this reason, the language model should be adapted dynamically [89].

In an ASR, dynamic language model adaptation is a common strategy to decrease the WER of the transcription given by the ASR by providing language models with a higher expectation of words and word-sequences that are typically found in the topic or topics of the story that is being analyzed. This method has shown to be effective in tasks that comprise a large amount of documents on different topics and also for processing data from multi-domain applications such as broadcast news transcription [90, 91].

Echeverry et al. in [92] proposed a dynamic language model adaptation in order to improve the accuracy of ASR in two phases. They used a linear interpolation between a background general language model and a topic dependent language model.

Language model adaptation approaches can be categorized according to different criteria. Rosenfeld in [93] suggested a classification based in the domain of the data. On the contrary, Bellegarda in [94] proposed that the classification must be done according to the system requirements. However, there is not a distinct separation between these criteria. Nowadays language model adaptation approaches jointly depend not only on the origin and domain of the data but also on the system requirements and the objective of the adaptation scheme. Some language model adaptation techniques depend on the specific context of the task that they are addressing. In these techniques, new sources of information are used to generate a context-dependent language model which is then merged with a static language model. For example, these new sources of information may come from text categorization systems as in [95], from speaker identification

systems in [96], from linguistic analysis systems in [97] or from the application context itself in [98].

Additional techniques depend on analysis and extraction of metadata, which means extraction of information that it is not clearly described in the text. The topic of a document or semantic information related to it are examples of metadata. Latent Semantic Analysis is an instance of the kind of techniques that exploits this kind of information. Bellegarda in [94] showed that the use of LSA is proposed to extract the semantic relationships between the terms that appear in a document and the document itself. More robust techniques in the field of information retrieval, as Latent Dirichlet Allocation in [99], have also been used for adapting language models in the ASR task in [100]. A keyword extraction strategy to determine the language model to be used in a multi-stage ASR system is proposed in [101]. In contrast to Latent Semantic Analysis, which does not explicitly consider the exact word order in the history context, in [102] a history weighting function is used to model the change in word history during LM adaptation.

However, in this study the researcher used the language model interpolation approach that is a simple and widely used method for combining and adapting Arabic language models. It consists of taking a weighted sum of the probabilities given by the component models.

3.3 Pronunciation Model Adaptation Approaches

The main problem faces ASR applications is variability in speech. Data-driven and Knowledge-based are two methods for modeling variation in speech or pronunciation variations [67]. Data-driven methods based on the training corpus to find possibly of the pronunciation variants (direct data-driven) or transformation rules (indirect data-driven). Although the indirect data-driven method generates rules that are used to find variants, the direct data-driven method generates only variants. On the other hand, the knowledge-based methods are based on linguistic standards. These standards are presented as phonetic rules that can be employed to find possible alternatives of pronunciation for word utterances. However, the knowledge-based method cannot describe all variations that occur in continuous speech, while obtaining effective information using the data-

driven method is very difficult [103]. The Knowledge-based model needs domains' expertise to generate phonetic rules [104].

Ali et al. in [105] presented a technique that is a rule to generate phonetic dictionaries for a large vocabulary of Arabic speech recognition system. They consider MSA as well as some common dialectal cases in generating the pronunciation rules. Kirchhoff et al. in [106] suggested using Romanization approach for transcription of Egyptian dialectic in telephone conversations. Ali et al. in [107] suggested a rule-based approach to generate Arabic phonetic dictionaries for a large vocabulary speech recognition system. The system used classical Arabic pronunciation rules, common pronunciation rules of Modern Standard Arabic, as well as morphologically driven rules.

Strik in [108] showed that pronunciation variations modeling should be considered at the three mentioned levels. Lexical adaptation is the traditional method that is adding variants to the pronunciation dictionary. Sloboda and Waibel in [109] presented that the key of improving the accuracy in continuous ASRs is adding alternative pronunciations in dictionary. Fosler-Lussier et al. in [110] demonstrated that the WER increased when there is a mismatch between the phones recognized and the word's phonetic transcription in the dictionary.

Alghamdi et al. in [64] generated MSA news broadcasting transcription application based on rule-based approach. Al-Haj et al. in [111] showed that the linguistic pronunciation rules were used for generating variants in dictionary from Iraqi-Arabic speech. Biadsy et al. in [112] showed that using the knowledge-based approach can improve phone recognition and word recognition outcomes. They generated a set of pronunciation rules that cover within-word variation in MSA speech.

AbuZeina et al. in [53] proposed direct data-driven approach to model within-word pronunciation variations. They proposed a method to extract pronunciation variants from the training news broadcasting corpus. AbuZeina et al. in [24] demonstrated that the knowledge-based approach which they applied to model cross-word pronunciation variation at phonetic dictionary and language model levels was effective in overcoming some problems .

Ramsay et al. in [113] proposed method for automatically generating a phonetic sequence for fully diacriticised Arabic text which closely matches the Arabic

pronunciation. They used a set of phonological rules that work on fully converting text into the actual sounds. Hyassat and Abu Zitar in [25] presented an Arabic speech recognition system using Sphinx-4[2]. They built pronunciation dictionaries for the Holy Qur'an and standard Arabic language using an automatic toolkit. In addition, they developed three corpuses which are the Holy Qura'an corpus of about 18.5 h, the command and control corpus of about 1.5 h, and the Arabic digits corpus of less than 1 h of speech.

3.4 An Arabic Speech Recognition

Many researchers proposed to develop automatic Arabic speech recognition system. For example, Al-Otaibi in [114] proposed various approaches for building a digital corpus of the Arabic speech. The author provides a new technique for labeling Arabic speech. The proposed technique was built using Hidden Markov Model (HMM) toolkit (HTK). As mentioned in [114], the proposed technique achieved a recognition ratio for speaker dependent ASR of 93.78%. Hyassat and Abu Zitar in [25] presented an Arabic speech recognition system using Sphinx-4. The authors built pronunciation dictionaries for the Holy Qur'an and standard Arabic language using an automatic toolkit. In addition, they developed three corpuses which are the Holy Qura'an corpus of about 18.5 hours, the command and control corpus of about 1.5 hours, and the Arabic digits corpus of less than 1 hours of speech. In [25, 114] the authors proposed to develop Arabic ASR by building Arabic speech corpus, However, building such a corpus considered time efforts and time consuming, due to the lack of spoken and written training data contradicted by Arabic ASR.

Ali et al. in [105] presented a technique that is a rule to generate phonetic dictionaries for a large vocabulary of Arabic speech recognition system. They considered MSA as well as some common dialectal cases in generating the pronunciation rules. The WER came to 9.0% using the proposed method. Kirchhoff et al. in [106] suggested using Romanization approach for transcription of Egyptian dialectic in telephone conversations. AbuZeina et al. in [24] demonstrated the cross-word

[2] Speech recognition system has been jointly developed by Carnegie Mellon University, Sun Microsystems Laboratories, and Mitsubishi Electric Research Laboratories (MERL)

pronunciation variation for continuous Arabic speech recognition. They presented a knowledge-based approach to model cross-word pronunciation variation at both phonetic dictionary and language model levels. The enhanced method achieved WER of 9.91% on a fully discretized transcription of Arabic broadcast news. AbuZeina et al. in [53] proposed direct data-driven approach to model within-word pronunciation variations. They proposed method to extract pronunciation variants from the training speech corpus. The enhanced method achieved WER of 11.17% when the variants are represented within the language model. However, the proposed method did not add considerable improvements by expanding dictionary alone.

Ali et al. in [107] suggested a rule-based approach to generate Arabic phonetic dictionaries for a large vocabulary of speech recognition system. The system used classical Arabic pronunciation rules, common pronunciation rules of Modern Standard Arabic, as well as morphologically driven rules. The proposed approach achieved WER of %11.71 for fully diacritized transcription of Arabic broadcast news. Ramsaya et al. in [113] proposed a method for automatically generating a phonetic sequence for fully diacriticised Arabic text which closely matches the Arabic pronunciation. They used a set of (language-dependent) pronunciation rules that work on fully converting a text into the actual sounds. Abdou et al. in [115] showed a speech-enabled computer-aided pronunciation learning (CAPL) system. The system was developed for non-native speakers to teach the Arabic pronunciation. The system also uses a voice recognition unit to detect errors in user recitation. A phoneme duration of the algorithm was implemented in order to detect recitation errors related to phoneme durations. Accuracy evaluation using a dataset that includes 6.6% wrong speech segments reported that the system correctly identified the error in 62.4% of pronunciation errors, reported “Repeat Request” for 22.4% of the errors, and made false acceptance of 14.9% of total errors. The studies in [24, 53, 105-107, 113, 115] proposed to use the pronunciation rules in order to build a dictionary model that spells Arabic words based on Arabic pronunciation rules. Due to shortage of spoken and written training data to build an Arabic corpus for ASR, they stressed the importance of distilling pronunciation variants from training speech corpus. Hence, the authors mentioned that this does not cover all pronunciation rules.

Soltau et al. in [116] presented evolution in Arabic speech recognition system in the IBM as part of the continuous effort for the Global autonomous language exploitation (GALE) project. The system consists of different stages that incorporate both vocalized and non-vocalized Arabic speech model. The system also incorporated a training corpus of 1,800 hours of unsupervised Arabic speech. Azmi et al. in [117] suggested Arabic syllables for speaker-independent speech recognition system for Arabic spoken digits. There are 44 Egyptian speakers for both training and testing in their corpus. The experiments in a clean environment, reported that the recognition rate obtained using syllables perform better than the rate obtained using monophones, triphones, and words by 2.68%, 1.19%, and 1.79%, respectively. Also in noisy telephone channel, syllables perform better than the rate obtained using monophones, triphones, and words by 2.09%, 1.5%, and 0.9%, respectively.

Khasawneh et al. in [3] compared a polynomial classification that was utilized to isolated-word speaker-independent Arabic speech and dynamic time warping (DTW) recognizer. They reported that the polynomial classification produce better recognition accuracy, and faster response than DTW recognizer does. Rambow et al. in [118] showed the problem of parsing transcribed spoken Arabic. They studied three different methods: sentence transduction, treebank transduction, and grammar transduction. They reported that grammar transduction perform better than the other two approaches do.

Nofal et al. in [119] presented implementation of stochastic-based new acoustic models adequate for use with a command and control system speech recognition system for Arabic language. However, the study does not consider continuous Arabic speech and large vocabulary speaker-independent for speech recognition system.

Several researchers have suggested the use of neural networks for Arabic phonemes and digits recognition [120, 121]. Alimi and Ben Jemaa in [120] presented the use of a fuzzy neural network for isolated words recognition. El-Ramly et al. in [121] studied that the use of artificial neural networks (ANN) in recognition of Arabic phonemes. Bahi and Sellami in [122] studied a hybrid of neural network and hidden Markov model NN/HMM for speech recognition. However, the coverage of any corpora cannot contain complete information about all aspects of language lexicon and grammar [3], due to the limited written training data and therefore inadequate spoken training data. So, the

neural networks can never train all phones in Arabic language. The NN requires time more than HMM technique.

All the previous studies tried to enhance the Arabic ASR. However, the main point in developing an Arabic ASR is the corpuses. The aforementioned studies show that there is a shortage of spoken data as compared to written data, resulting in a great need for more speech corpora in order to serve different domains of Arabic ASR.

CHAPTER 4

Methodology

This chapter presents the proposed methods for improving the accuracy of Arabic ASR system. The main idea of new approaches is that they adapt target acoustic model using appropriate acoustic models and create a suitable language model that has the new changes in the language that are not found in the current language model. The chapter discusses every step in detail.

4.1 Research Design

In this study eleven steps are followed as demonstrated in **Figure 4-1**.

1. The first step is acquiring of an Arabic speech corpus.
2. The second step of the research is data preparation.
   2.1. Generating control file (fileids).
   2.2. Generating transcription files.
   2.3. Generating filler file.
   2.4. Generating phone file.
3. The third step is generating pronunciation model.
4. The fourth step is generating language model.
5. The fifth step of the research is generating the acoustic model.
   5.1. Splitting the dataset to training part and testing part.
   5.2. Training the data (Trainer).
6. Estimating the WER of the baseline of Arabic ASR system.
7. Generating another Arabic speech corpus.
   At this step, we will use the new corpus to generate another acoustic model (source acoustic model) that employed within acoustic model approaches.
8. Designing and developing adaptation approaches.
   8.1. Acoustic model adaptation approaches:
      8.1.1. Interpolation acoustic model approach.

8.1.2. Merging acoustic model approach.

8.1.3. Hybrid of interpolation and merging of acoustic model approach.

8.1.4. Enhancing hybrid approach.

8.2. Language model adaptation approach:

8.2.1. Interpolation language approach.

8.3. Pronunciation approaches:

8.3.1. Removing all diacritized text.

8.3.2. Eliminating all duplicate in pronunciation of the word.

8.3.3. Adding Al-Shamsi and Al-Moon.

8.3.4. Replacing FATHA followed by WAW to WAW.

8.3.5. Split WAW rule.

8.3.6. Unifying pronunciation of Tanween.

8.3.7. Merging pronunciation of FATHA, Long FATHA, Pharyngeal Version of FATHA, and Long Version of Pharyngeal Version of FATHA.

8.3.8. Converting pronunciation of Pharyngeal Version of DAMMA to DAMMA.

8.3.9. Converting pronunciation of Pharyngeal Version of KASRA to KASRA.

9. Decoding the data (Decoder).

We used Spinx3 decoder engine with testing data to generate transcription file. The decoding phase uses the trained acoustic model, pronunciation model, and laguage model.

10. Estimating the accuracy of new model using the WER.

10.1. After each experiment we require to estimate the accuracy using the WER.

10.2. Observing and measuring how well the methods support a solution to the problem.

11. Comparing with previous work.

11.1. Comparison between a new values of WER with the baseline of the WER.

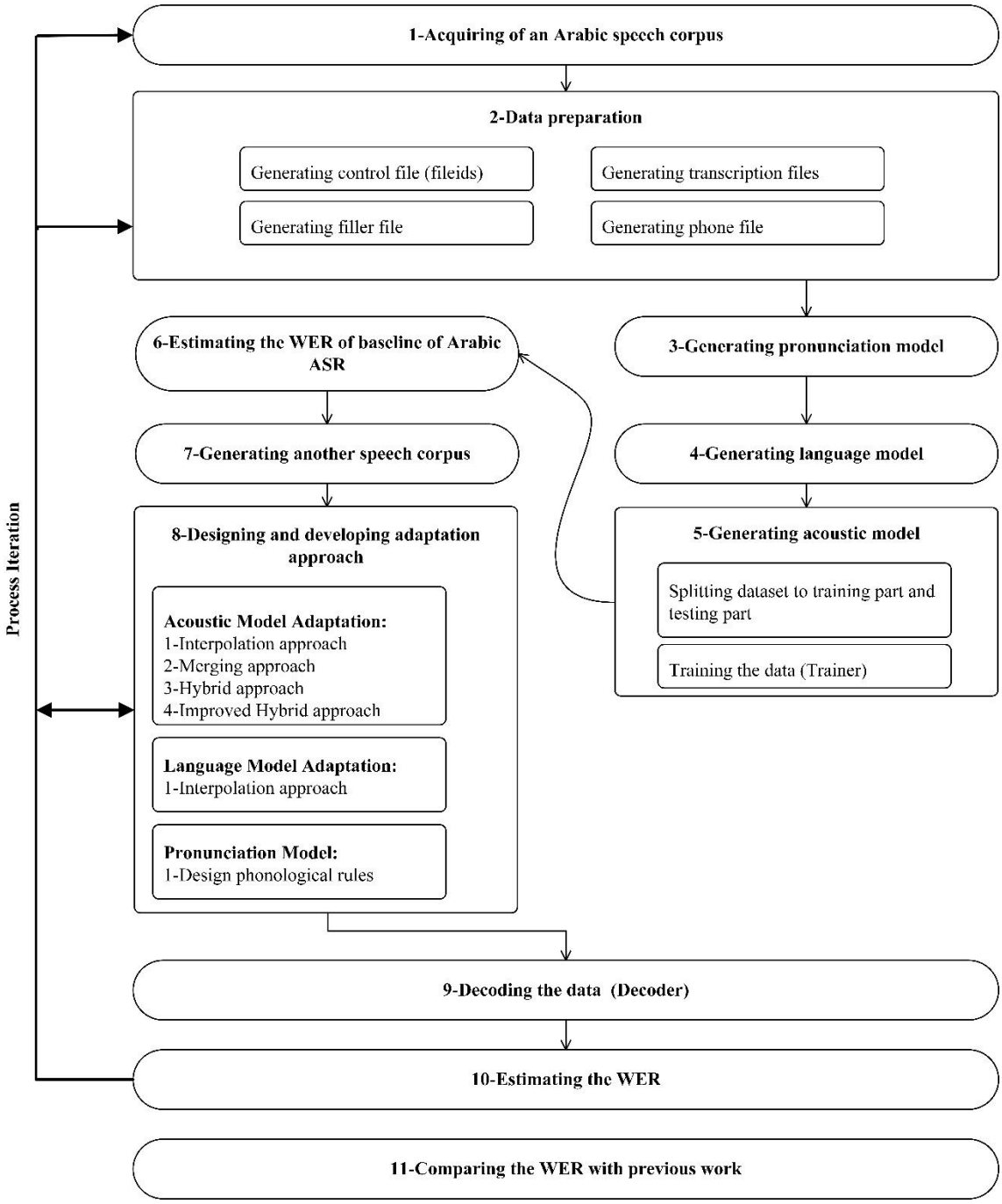

Figure 4-1 Methodology flowchart

To implement and evaluate the proposed methods, we needed to follow the processes as following:

4.2 Acquiring and Preprocessing Data

Stage 1. Acquiring of an Arabic speech corpus

In order to use a speech recognition system, we need data. This data is from speech that has been recorded. Speech data is needed both for training and for testing. The Arabic broadcast news corpus in [64] is from King Fahd University of Petroleum and Minerals was selected for examining the our methods in this study. Moreover, the broadcast news (BCN) corpus used as baseline to our experiments. This recognizer was designed to recognize continuously spoken Arabic broadcast news. In chapter 5 we will see more details about the corpus.

Stage 2. Building pronunciation rules experiment

In order to build a pronunciation model in an ASR system, we need rules to generate phonetic transcription for all vocabulary in this research. We selected a software tool that was developed in [24, 53, 105] for generating the Arabic lexicon. This tool is rule-based method to automatically generate an Arabic lexicon for a given diacritized transcription.

In addition, this study used the closed vocabulary to build the pronunciation dictionary. Closed vocabulary is a list of all transcription of words in the dictionary.

Stage 3. Baseline experiment

The aim of this experiment was to determine the WER as a bench mark for other next experiments. This experiment was performed on Arabic broadcast news corpus. In order to record WER as baseline in this research, we need to rebuild acoustic model, language model, and pronunciation model. The steps were as follows:

1. Data preparation

To prepare the data for experiments, we adopted the following steps:

1.1. Generating control file (fileids)

A control file (fileids) contains the list of filenames. It should not contain the extension (.wav). This step generated three files which are fileids, train.fileids, and test.fileids files. An example of the entries is given in **Figure 4-2**.

```
909/909-01
910/910-01
911/911-01
```

Figure 4-2 Example of fileids file from BCN

1.2. Generating transcription files

A transcription file is that transcripts corresponding to the wav files are listed in exactly the same order as the feature filenames in control file (fileids). This step was applied for training and testing transcription files. An example of the entries in **Figure 4-3**.

```
<s> أَمَّا فِي الأُرْدُن فَقَد تَمَّ وَضعُ بَرنَامِج ضَخم لِتَطوِير مَدِينَةِ العَقَبَة </s> (909-01)
<s> وَذَلِكَ بِصَالَةِ نَادِي المُعَاقِين الكُوَيتِيِّ </s> (910-01)
<s> بِأُحدَثِ صُورَةٍ تِقنِيَّة </s> (911-01)
```

Figure 4-3 Example of transcription file from BCN

1.3. Generating filler file

A filler dictionary contains the non-speech events (not covered by language model or non-linguistic) such as breath, hmm or laugh and should be mapped to user defined phones. An example of the entries is in **Figure 4-4**.

```
<s>        SIL
</s>       SIL
!INH       +INH+
!NOISE     +NOISE+
```

Figure 4-4 Example of filler file

Where the first column represents the non-speech events or noises that exist in wave file in speech corpus. The second column presents phones defined by the user for each noise in first column. Note that the words <s> and </s> are treated as special words and are required to be presented in the filler dictionary. At least one of these must be mapped on to a phone called "SIL". The phone SIL is treated in a special manner and is required to be presented. The <s> is beginning-utterance silence and </s> is end-utterance silence.

1.4. Generating phone file

A phone file is a list of all phones in the Arabic language, in addition to all fillers. An example of the entries is in **Figure 4-5**.

```
SIL
+INH+
+NOISE+
AE
AA
AA:
```

Figure 4-5 Example of phone file

List of phones is a list of all acoustic units that you want to train models. The SPHINX-3 does not permit to have units other than those in the dictionaries. All units in your two dictionaries must be listed here. In other words, the list of phones must have exactly the same units used in the dictionaries.

2. Generating language model (The task grammar)

We used CMU-Cambridge Statistical Language Modeling (SLM) toolkit [123]. The SLM is a suitable software tools to facilitate the construction and testing of statistical language models. It is used to build a statistical language model from the transcription of the full diacritized transcription of the Arabic broadcast news speech. The language model was built using 3-grams. **Figure 2-5** shows that the process for creating a language model. The process for creating and testing language model are as follows:

1. Generating the word frequency (unigram) file from the Arabic broadcast news transcription (train and test transcription). The output is that list of every word which occurred in the text, along with its number of occurrences.

2. Generating the vocabulary file. The output is the file containing a list of vocabulary words.
3. Generating the n-gram (bi-grams and tri-grams) file. The output is the list of every id n-gram which occurred in the text, along with its number of occurrences.
4. Generating the ARPA file. The output is a language model, in binary format.
5. Converting ARPA file to DMP file. The output is that a DMP file, in binary format to use within Sphinx-3.

3. Generating pronunciation dictionary (Pronunciation model)

We used the java tool, proposed in stage 2, to generate the pronunciation model using diacritized transcription of the broadcast news corpus. An example of the entries in **Figure 4-6**.

أَثَرَ E AE TH AE R AA
أَثَرِ E AE TH AE R IX

Figure 4-6 Example of pronunciation dictionary file

The first column represents each Arabic diacritized word in text corpus. The second column represents possible sequence phones.

4. Training the data (Trainer)

The ASR system would typically look at the speech data and then use it to understand speech input. The phase where it looks like learning and is usually called the training phase.

We used SphinxTrain [124]. SphinxTrain software provides a set of tools to create acoustic models for ASR applications. The Sphinx trainer is based on HMM. An HMM-based system, like all other speech recognition systems, functions by first learning the characteristics (or parameters) of a set of sound units, and then using what it has learned about the units to find the most probable sequence of sound units for a given speech signal. The process of learning about the sound units is called training.

Once that is done, the Sphinx trainer generated four files which are mixture weights, transition matrices, means, and variances files that can be used by the decoder for speech

recognition. These four files provide context dependent of an acoustic model. Mixture weights file contains the weights given to every Gaussian in the Gaussian mixture corresponding to a state. Transition matrices file contains the matrix of state transition probabilities. Means file contains means of all Gaussians. Variances file is the variances of all Gaussians.

5. Decoding the data (Decoder)

We used Sphinx3 [125]. Sphinx3 is one of Carnegie Mellon University's open source large vocabulary, speaker-independent continuous speech recognition engine. During this step we actually tested our trained data against some speech. After training, it's mandatory to run the decoder to check training results using Spinx3 decoder. The Decoder takes a model, tests part of the database and reference transcriptions and estimates the quality (WER) of the model. During the decoding stage we used the language model with the description of the order of words in the language. In addition, the decoding stage uses the trained acoustic model and pronunciation model.

6. Evaluation

Recognizing the test data can evaluate the results using the Speech Recognition Scoring Toolkit (SCTK) [126].

Our goal is to decrease the WER. In order to evaluate the WER of our proposed method we selected the most common metric which is computed using Equation **(7)**.

Stage 4. Acquiring another speech corpus and creating source acoustic model experiment

The aim of this experiment was to create a corpus for adapting data and building source trained acoustic model. The steps were as follows:

1. Data Preparation

1.1. Collecting speech data

For studying Arabic speech recognition, we have collected an Arabic speech corpus from the Holy Qur'an (HQ). The Holy Qur'an speech was uttered by one Arabic speaker.

The Holy Qur'an wave files [127] were selected to build second corpus and then build second acoustic model for testing the proposed methods. The main reasons of selecting the Holy Qur'an is that has fully diarized transcription and it is available in several speech forms by different persons. Basically, we need to record a single audio file for each sentence in the adaptation corpus, naming the files according to the names listed in transcription and fileids files. In addition, we recorded at a sampling rate of 16 kHz in mono with single channel and files format is ".wav". Each wave file represents only one Ayah from the Holy Qur'an. We selected 1230 wave file after transcription file in an ascending ordered according to length of transcription file.

1.2. Generating features

We extracted sequences of feature vectors from the raw speech waveforms. MFCCs were used in this work. This can be done with the wave2feat tool from SphinxBase.

In this step, the front-end module was used to pre-process the raw speech at 16 bits sample with sampling frequency of 16 kHz to cepstral feature vectors together with its first and second derivative. This produces feature vectors with a total of 39 dimensions. SphinxTrain makes use of the feature vectors to create a continuous HMM acoustic model. A phoneme or phone was used as the unit of HMM, and each has three states, with a left-to-right topology.

1.3. Generating control file (fileids)

A control file (fileids) contains the list of filenames. It should not contain the extension (.wav). This step generated two files which are fileids, and train.fileids files. An example of the entries in **Figure 4-7**.

```
Alafasy_6080
Alafasy_4902
Alafasy_5324
```

Figure 4-7 Example of fileids file from HQ

After that, we can repeat the same steps in baseline experiment to generate transcription file, filler file, phone file, and pronunciation dictionary from the HQ data. In addition, we can train and decode data against the HQ speech data by the same way in

baseline experiment. Finally, we evaluate the results using the same tool in the baseline experiment which is SCTK.

Stage 5. Experiment of cross-validation test

This research employed cross-validation dataset (5 fold) for training and testing purpose. Cross-validation attempts to mimic test data with these steps as follows: first, the sample is randomly divided into m equal-sized, non-overlapping subsamples (5 is a frequent default) [128]. Second, each of the m subsamples is for training while a testing from the other m - 1 subsamples. Third, the predictive accuracy of each WER is tested with the one subsample not included in its training set. Finally, averaging over the m proved to fit the assessment. This new method can be used to determine a reasonable value for the WER. So, the cross-validation can be used as a validation test.

In addition, we split the broadcast news corpus to 5 fold for testing and evaluation purpose. **Table 4-1** shows that summary of five subsamples from the Arabic broadcast news corpus that will use for generated several training and testing dataset.

Table 4-1 Summary of subsample from the Arabic broadcast news corpus used for generated several training and testing dataset.

Corpus	Words	Vocabulary	Speakers	Hours
Subsample 1	8,107	4,266	153	1.9
Subsample 2	7,894	4,275	187	1.8
Subsample 3	7,979	4,346	166	1.8
Subsample 4	7,982	4,316	130	1.8
Subsample 5	7,411	4,112	74	1.4

4.3 Experiments of Acoustic Modeling

Stage 6. Experiment of interpolation acoustic model.

The goal of the interpolation experiment is to adapt target acoustic model (Broadcast News) using source acoustic model (the Holy Qur'an). Interpolation acoustic

model can be achieved by estimating a priori weights to multiply them to the acoustic model of the target and source language. For simplicity, the target language (Arabic Broadcast news) is the language for recognition by Arabic ASR system (the spoken language), while the source language (The Arabic Holy Qur'an) is the language used for adjusting the target model. Euclidean distance measures certain distance between two points. Euclidean distance is used to select the nearest Gaussian in the source acoustic model to certain Gaussian in the target acoustic model as shown in **Figure 4-8**.

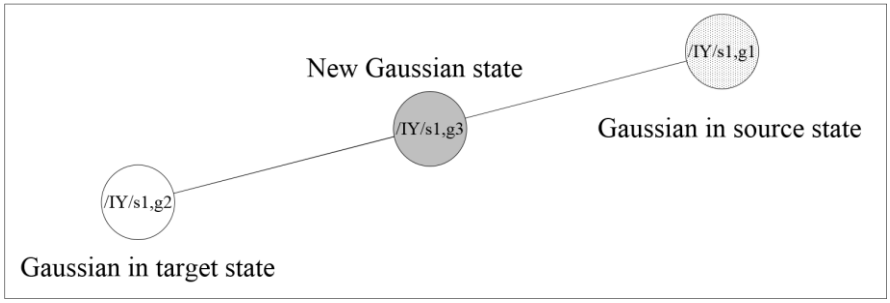

Figure 4-8 Acoustic space. Interpolation of target state (Arabic Broadcast news) and source state (Holy Qur'an) by setting weight at 0.5. The filled circle is the new created Gaussian.

We proposed to do this by first mapping each phoneme in the pronunciation dictionary of the source language to the phoneme of the target language based on phoneme (phonetic) knowledge as shown in **Figure 4-9**. Note that in our case one source phoneme should be mapped to the same target phoneme. It is a one-to-one mapping.

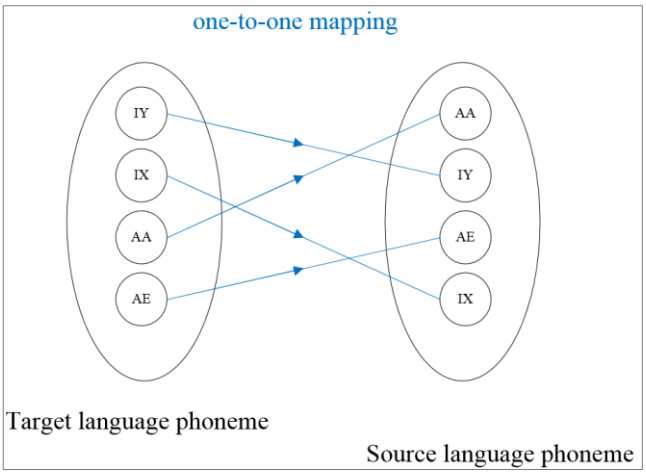

Figure 4-9 Phoneme mapping from target to source language

Second, we adapted the initial target language acoustic model using the source language corpus and pronunciation dictionary, instead of using Baum Welch algorithm to re-create the Gaussians. The resulting acoustic model has the same number of Gaussians as the target language acoustic model. A priori weights for each model are predicted, and a new model is created with equation (9) [80]:

$$AM_{new} = (1 - \beta).AM_{trg=BCN} + \beta.AM_{src=HQ}, 0 \leq \beta \leq 1 \qquad (9)$$

Where AM_{new} is the adapted acoustic model, AM_{trg} is the target acoustic model, and AM_{src} is the source acoustic model. The β is interpolation weight.

Stage 7. Experiment of merging acoustic model.

The goal of the merging experiment was also based on the idea of combining the model sets of source and target language of a formal Arabic speaker.

The requirements of this algorithm are a speaker dependent model set of the source language and a mapping between the two languages. The mapping model is shown in **Figure 4-9**. In this method, the Gaussian mixture of each state of the target language is merged with a state of the corresponding model of the source language according to the given mapping as shown in **Figure 4-10**. This merging of two Gaussian mixtures yields a new mixture with twice as many components as in the original mixtures.

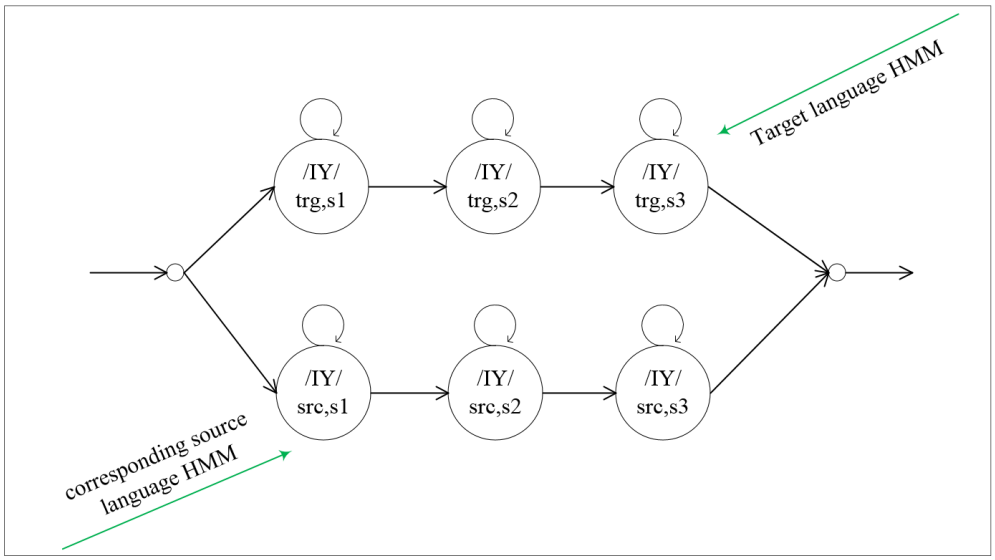

Figure 4-10 Acoustic model merging corresponding models from a source and a target acoustic.

The size of a Gaussian mixture associated with a state is expanded by copying mixture components from other mixtures. The choice of which mixture components to use for expanding is based on a criterion of minimized frame-level errors. Finally, the weights of all mixture components are re-estimated with Baum-Welch re-training.

In addition, two model sets were combined to form a new model set by merging Gaussian mixtures. In our case, the two model sets which were used for merging had been trained for the same language. The difference between the two model sets was that each set was trained with a different phone-level transcription of the training material.

Stage 8. Experiment of Hybrid of Acoustic Model Interpolation and Merging Approach for offline adaptation.

The goal of the hybrid experiment was to combine between interpolation and merging approaches. Acoustic model interpolation and merging approach for offline adaptation are a promising approaches to create a model which is intermediate between the target and source acoustic models. Therefore, an approach that integrates interpolation and merging seems suitable.

In the combination process, when the distance between a Gaussian in certain state of the target model (referred to as target Gaussian) and the corresponding Gaussian in certain state of the source model (referred to as source Gaussian) is below a threshold, their values will be interpolated. Otherwise, merging is performed between target Gaussian and source Gaussian.

The general approach of interpolation is to select the nearest Gaussian from the corresponding source state for every Gaussian in the target state using certain distance measure. Instead, we propose to execute the interpolation in a different way, where every Gaussian in the target state is treated like the "centroid" for the Gaussians in the source state. In order to find the nearest target Gaussian, at the next step, for all source Gaussians the researcher used distance measure like Euclidean distance **Equation (10)** [129].

$$Euclidean\ Distance\ = \sqrt{\sum(\mu_{tg} - \mu_{src})^2} \qquad (10)$$

where μ_{tg} and μ_{src} are two points from target Gaussians and source Gaussians respectively. Every source Gaussian will be associated with only one target Gaussian. So, certain target Gaussians are associated with zero or more source Gaussians. When the distance between the associated target Gaussian and the source Gaussian is below a threshold, their means, variances and mixture weights will be interpolated using the formula in **Equation** (11) [81]. On the other hand, merging is performed: for those target Gaussians without any associated source Gaussian using the formula in **Equation** (13) [81] or for the source Gaussian that are faraway (more than the threshold) from their associated target Gaussians using the formula in **Equation** (12) [81]. In **Equation** (12) and (13), their mixture weights will be reduced by the interpolation weight. The resulting model is a hybrid model of interpolation and merging.

The $g_{Src,m}$ is the set of source Gaussian associate with $g_{tg,n}$ the target Gaussian. The $g_{new,i}$ is the adapted model, while *d(.)* is the distance function and ω is the mixture weight. Also, ω_{new} is the adapted mixture weight and ω_{src} is the source mixture weight.

The β is interpolation weight. The *tg* is the broadcast news and the *src* is the Holy Qur'an. Also, *dis* is the threshold.

$$g_{new,i} = (1-\beta).g_{tg,n} + \beta.g_{src,m}, g_{src} \neq \emptyset, d(g_{tg,n}, g_{src,m}) \leq dis \quad (11)$$

$$g_{new,i} = g_{src,m}, \omega_{new,k} = \beta.\omega_{src,m}, g_{src} \neq \emptyset, d(g_{tg,n}, g_{src,m}) > dis \quad (12)$$

$$g_{new,i} = g_{tg,n}, \omega_{new,i} = (1-\beta).\omega_{tg,n}, g_{src} = \emptyset \quad (13)$$

The target language in this case is the new acquired language of the Arabic speakers. The possible source languages can be any of the three types of languages we mentioned in the earlier section (L1, L2 and L3). The source language that can be used in our case is L1 (native Arabic language).

The target and source acoustic models may have different configuration in terms of number of states and number of Gaussians. Therefore, before the modeling can be carried out, the states and Gaussians of the target and source acoustic model have to be matched. In our current implementation, we use a simple context matching. This means that in cases where the models used for modeling are CD models, the matching triphone in the source model will be looked upon.

Stage 9. Experiment of validating the hybrid method.

In this experiment, we need to verify the accuracy in previous experiment (Hybrid of Acoustic Model Interpolation and Merging Approach for offline adaptation) by using the cross-validation evaluation. Also, it is difficult to compare accuracy of WER produced and requires a cross-validation method to avoid different WER outcomes because of different initial partitions of dataset.

To make certain results of the WER, 5-fold cross validation was used to find a generalized, and optimal of the WER. In addition, we estimated the WER for five times on 5 datasets across every weight. Then, we averaged the WER for each weight.

Stage 10. Experiment of several distance equations within hybrid method.

In this experiment, we applied several distance equations to estimate the distance between sources and target Gaussian. **Table 4-2** shows varied distance equations proposed to apply within hybrid approach. Subsequently, we could estimate the best distance equations within hybrid approach, where the distance is estimated between two points, which are *p* and *q*, in space. In our case, *p* and *q* represent target Gaussians and source Gaussians.

Table 4-2 Summary of Several Distance Equations Used within Hybrid Approach

Eq.#	Distance equation	Named	Reference
1.	$d(p,q) = \sqrt{\sum (p-q)^2}$	Euclidean distance	[129]
2.	$d(p,q) = \sum \lvert\lvert p-q \rvert\rvert$	Manhattan distance	[130]
3.	$d^2(p,q) = \sum (p-q)^2$	Squared Euclidean distance	[129]
4.	$d(p,q) = \sqrt{\sum p^2 + q^2 - 2pq}$	Equivalent to Euclidean distance	[129]
5.	$d(p,q) = \sqrt{\sum (p^2 + q^2)}$	Pythagorean distance	[131]
6.	$d(p,q) = \sqrt[2]{\sum (p-q)^2}$	Minkowski distance (p=2)	[132]
7.	$d(p,q) = \sqrt[3]{\sum (p-q)^3}$	Minkowski distance (p=3)	[132]
8.	$d(p,q) = \sqrt[4]{\sum (p-q)^4}$	Minkowski distance (p=4)	[132]

Stage 11. Experiment of evaluating the best result from several distance equations.

In this experiment, we validated the best distance equations, according to the best result recorded in the previous experiment (several distance equations within hybrid method). The aim of our experiment is to verify that these results are not just due to

some peculiarities of the Arabic speech, the same experiments are redone with different training and testing datasets speech. This validation was applied on 5-cross-validation.

4.4 The Experiments of Language Modeling

Stage 12. Experiment of interpolation language model.

Language model interpolation is a simple used method for combining and adapting language models. It consists of taking a weighted sum of the probabilities given by the component models, using the formula in equation (14) [92]:

$$LM_{new} = (1 - \lambda).LM_{trg} + \lambda.LM_{src} \quad (14)$$

where λ is the interpolation weight between both models, which has to fulfill the condition $0 \leq \lambda \leq 4$.

In general, we smoothed the broadcast news model using the information from the Holy Qur'an model. The process for the generation of the LM_{new} in this work was followed in the different stages as follows:

1. We replicated transcription of the broadcast news four times (target language) in order to generate the new language model.
2. We replicated transcription of the broadcast news three times (target language) and the Holy Qur'an one time (source language) in order to generate the new language model.
3. We replicated transcription of the broadcast news two times (target language) and the Holy Qur'an two times (source language) in order to generate the new language model.
4. We replicated transcription of the broadcast news one time (target language) and the Holy Qur'an three times (source language) in order to generate the new language model.
5. We replicated transcription of the Holy Qur'an four times (source language) in order to generate the new language model.

Finally, it is generally difficult to examine the effect of language model directly. For most of the language model smoothing research, the accuracy is measured based on

extrinsic evaluations. We included extrinsic evaluations in this study, i.e., the WER for all recognition system.

Stage 13. The Experiment of combining interpolation language model followed by The Hybrid acoustic model.

In this experiment, we proposed the combination of the two specific approaches which are interpolation language model and Hybrid acoustic model approaches. We aimed to improve the WER of Arabic ASR system in our study as much as possible. This combination approach consists of the following steps:

1. Building the trained acoustic model based on the broadcast news speech (target).
2. Building the trained acoustic model based on the Holy Qur'an speech (source).
3. Selecting the best language model based on the WER that was recorded from interpolation language model experiment.
4. Applying hybrid of acoustic model interpolation and merging approaches for offline adaptation.
5. Evaluating the new ASR system.

4.5 The Experiments of Pronunciation Modeling

Stage 14. Experiment of rules in pronunciation model.

The knowledge-based approach and the data-driven approach are two approaches pertaining to pronunciation model adaptations for ASR. The knowledge-based approach provides pronunciation rules from phonological knowledge and develops a pronunciation dictionary based on the pronunciation rules. However, the data-driven approach, phonological rules for pronunciation adaptation are automatically generated from speech and transcription data; as such, a subdivision into direct and indirect data-driven methods can be applied.

Pronunciation rules were used to transform a base form into a pronunciation variant. The phonological rules were derived based on linguistic and phonological knowledge according to known pronunciation variations of speech. Then, the phonological rules

were applied to base forms in a pronunciation dictionary. We selected the indirect data-driven approach to generate pronunciation rules in this study. We generated several pronunciation rules such as:

1. Removing all diacritized text.
2. Eliminating all duplicate in pronounce the word.
3. Adding Al-Shamsi and Al-Moon.
4. Replacing FATHA followed by WAW to WAW.
5. Splitting WAW rule.
6. Unifying pronunciation of Tanween.
7. Merging the pronunciations of FATHA, Long FATHA, Pharyngeal Version of FATHA, and Long Version of Pharyngeal Version of FATHA.
8. Converting pronunciation of Pharyngeal Version of DAMMA to DAMMA.
9. Converting pronunciation of Pharyngeal Version of KASRA to KASRA.

CHAPTER 5

Experiments and Evaluation

This chapter shows the experiments that have been conducted to examine the proposed methods. In the next section, an overview of the experimental setup will be given. The experiments are divided into three parts. The first part consists of the experiments which examine the acoustic modeling approach. The second part contains the experiment which tests on language modeling, followed by the third part which tests pronunciation modeling.

5.1 The Experimental Setup

First section provides an overview of the speech recognition system utilized for speech recognition. Then, we present an overview of the corpora used for training, and testing throughout our experiments.

5.1.1 Experimental Environment

Our experiment used HP laptop ProBook with an Intel Core i5-3230M, 2.60GHz processor with 6GB of RAM, and a 1T hard drive. The operating system is Ubuntu 13.10 64-bit.

5.1.2 Automatic Speech Recognizer: Sphinx-3

All experiments of this work have been performed using Carnegie Mellon University (CMU) Sphinx-3 [38] for the speech recognition tasks. Sphinx is an open source toolkit for speaker-independent, large vocabulary, and continuous speech recognition. Sphinx-3 can be divided into two main systems. The two main systems are the trainer and the decoder. The trainer, known as SphinxTrain, used for training continuous HMM acoustic models. CD modeling is part of the SphinxTrain application. The iterative re-estimation procedure is employed by using Baum Welch algorithm. Moreover, for creating robust triphone context dependent models, states are being tied together, that are called senones, during the training process. When multiple HMM states

share the same Gaussian mixture, it said to be shared or tied. These shared states are called tied states (known as senones). Additionally, Sphinx-3 is the application for decoding speech. Sphinx-3 is a fast decoder, capable of decoding speech at real time. This is achieved using conventional Viterbi search strategy and beam heuristics. Moreover, it has a lexicon-tree search structure. Sphinx-3 uses the acoustic model created by SphinxTrain. For language model, it accepts n-gram model in binary format, which is converted from a standard ARPA n-gram model. The Sphinx speech recognition package provides two major ways to do speaker adaptations. One is called MAP re-estimation of the parameters. The other is Maximum Likelihood Linear Regression (MLLR).

Finally, in the following experiments are based on the Sphinx ASR toolkit. In all experiments conducted in this thesis, the front-end module was used to pre-process the raw speech at 16 bits sample with sampling frequency of 16 kHz to cepstral feature vectors together. This produces feature vectors with a total of 39 dimensions. SphinxTrain makes use of the feature vectors to create a continuous HMM acoustic model. Phoneme or phone were used as the unit of HMM, and each has three states, with a left-to-right topology. On the contrary, the n-gram language model was created using CMU statistical language modeling toolkit

5.2 Speech Corpus

In this section, we summarize the main components of the baseline ASR system that is used to test the proposed method. Next, the Arabic speech recognition components contain the Arabic speech corpus, Arabic phoneme set, Arabic language model, and Arabic pronunciation dictionary. Moreover, all experiments were carried out on Arabic speaker-independent, large vocabulary, automatic, and continuous speech recognition. The following sections show an overview of the Broadcast News (BCN) and the Holy Qur'an (HQ) corpus used for training and testing the Arabic speakers. The broadcast news and the Holy Qur'an corpus were also employed for creating Arabic acoustic models, language model and pronunciation model.

5.2.1 Arabic broadcast news speech corpus

Testing the proposed method is performed on an actual Arabic Automatic Speech Recognition application. The speech corpus and baseline Arabic ASR systems were developed at King Fahd University of Petroleum and Minerals [64]. The Arabic ASR uses the large vocabulary, speaker independent, natural Arabic continuous speech recognition system. The corpus of the system was collected from several radio and TV news. The speech corpus contains sport news, and economic news. The Arabic ASR system was built using CMU Sphinx 3 systems. Three emitting states of Hidden Markov Models for triphone-based acoustic models were used in the ASR system. The ASR system is trained at a sampling rate of 16 kHz samples per seconds. A continuous density of 8 Gaussian mixture distributions provides in the state probability distribution.

Table 5-1 describes the Arabic broadcast news speech corpus employed for training and testing acoustic model. The Arabic broadcast news corpus contains 144 male speakers, and 105 female speakers. In addition, it contains 249 stories which are related to sport news, and economic news, and summing up to 5.4 hours of speech. The 1.1 hours of Arabic broadcast news was used for testing purpose [64].

All speech files are completely transcribed with fully diacritized text. Transcription in MSA. The transcription means that the way the speaker has uttered the words, even if the utterance is grammatically incorrect. The transcription files, which are train and test transcription files, contain 39,373 words (with within-word variants). A word is defined here to be any sequence of letters and diacritical marks delimited by the space character. The vocabulary list contains 14,234 words (without variants).

Table 5-1: Summary of the Arabic broadcast news corpus used for training and testing.

Task	Corpus	Words	Vocabulary	Speakers	Hours	WER (%)
Training	Arabic broadcast news	30,026	14,234	189	4.3	
Testing	Arabic broadcast news	9,347	14,234	90	1.1	12.4

Finally, the acoustic model, language model, and pronunciation model, within BCN corpus, are regenerated to get baseline of WER. The WER of Arabic BCN corpus is 12.4% using CMU sphinx 3 system.

There are some of the files which still contain some kind of noise. First, some files contain background noise such as low level or fainting music. Second, some files contain environmental noise such as a reporter in an open area, for example, a stadium or a stock market. Third, some files contain low level overlapping foreign speech, occurring when a reporter is translating foreign statements.

5.2.2 Cross validation dataset

Cross-validation attempts to mimic test data as follows: first, the sample is randomly split into m equal-sized, non-overlapping subsamples (5 is a frequent default). Second, each of the m subsamples is for training while a testing from the other m - 1 subsamples. Third, the predictive accuracy of each WER is tested with the one subsample not included in its training set. Finally, averaging over the m fit assessments.

Table 5-2 Summary of cross validation dataset from the Arabic broadcast news corpus used for training and testing.

Corpus	Task	Words	Vocabulary	Speakers	Hours	WER (%)
Dataset1	Training1	31,962	12,097	189	4.35	
	Testing1	7,411	4,112	74	1.4	13.8
Dataset2	Training2	31,391	12,012	189	4.35	
	Testing2	7,982	4,317	130	1.8	13.8
Dataset3	Training3	31,394	11,970	189	4.31	
	Testing3	7,979	4,347	166	1.8	14
Dataset4	Training4	31,479	12,068	189	4.31	
	Testing4	7,894	4,276	187	1.8	13.2
Dataset5	Training5	31,266	12,065	187	4.29	
	Testing5	8,107	4,267	153	1.9	11.6
Average of WER (%) for all testing data						13.28

Table 5-2 shows that an overview description of five datasets used for evaluating the proposed methods. The total average WER of 5 datasets is 13.28% using CMU sphinx 3 system. We used this average of WER as baseline for our study.

5.2.3 The Holy Qur'an speech corpus

Speech corpus is required for generating the source acoustic model. Then, the proposed approaches can be tested. For this purpose, HQ speech corpus is used. The second speech corpus is the Holy Qur'an. The HQ scripts can be formal and standard form of Arabic, classical Arabic. Because these scripts have full diacritical marks, the Arabic phonetics are completely represented.

For training purpose, we started by a fully diacritized transcript of The HQ [133]. The HQ's transcription file consists of 77,888 words (with within-word variants) and contains 17,600 words (without variants). The text is converted to a word list after removing numbers, special symbols, and the Arabic letter extension character. A word is defined here to be any sequence of letters and diacritical marks delimited by the space character.

In addition, for training purpose we created new corpus from BCN and part of the HQ speech. The Holy Qur'an speech was uttered by one Arabic speaker. We selected only 1230 speech files from the HQ speech after ordered ascending, then added to the BCN for creating new dataset speech.

Table 5-3 describes the Holy Qur'an speech corpus employed for adapting Arabic acoustic model. So, the new corpus, which is a combination between 5.4 hours of BCN speech and 1.57 hours of HQ speech, summing up to 6.15 hours for evaluating proposed method. The new corpus contains 190 speakers after we added a new male speaker to new corpus. The corpus is based on CMU Sphinx 3 ASR system. Three emitting states HMM for triphone-based acoustic models are used in the ASR system. A continuous density of 8 Gaussian mixture distributions provides in the state probability distribution. The audio files were sampled at 16 kHz. The ASR system was trained at a sampling rate of 16 k samples per seconds. The WER of the Holy Qur'an corpus is 13.6% using CMU sphinx 3.

Table 5-3: Summary of the Holy Qur'an corpus used for training.

Task	Corpus	Words	Vocabulary	Speakers	Hours	WER (%)
Training	The Holy Qur'an	77,888	17,600	190	6.15	13.6

5.2.4 Arabic phoneme set

A phoneme is the basic unit of speech used in ASR systems.

Table 5-4 shows the listing of the Arabic phoneme group, which is 42 phonemes, used in the training, and the corresponding phoneme symbols. This phoneme group is selected based on the previous research with Arabic text-to-Speech systems [53, 105, 134-136].

Table 5-4: Set of phoneme used in the training

Phoneme	Letter	Example	Phoneme	Letter	Example
/AE/	Fatha ـَ	بَ	/D/	(Dal) د	أَدَاء
/AE:/	ـَا	بَاب	/DH/	(Thal) ذ	أَخِذَ
/AA/	Hard Fatha ـَ	خَ	/R/	(Raa) ر	آبَارِ
/AA:/	ـَ	خَاب	/Z/	(Zain) ز	أَبرَز
/AH/	Soft Fatha ـَ	قَ	/S/	(Seen) س	أَحَسِبَ
/AH:/	ـَ	قَال	/SH/	(Sheen) ش	أَشخَاصٍ
/UH/	Damma ـُ	بُ	/SS/	(Sad) ص	أَصبَحَ
/UW/	وُ	دُون	/DD/	(Dad) ض	أَضعَافِ
/UX/	ـُ	غُصن	/TT/	(Taa) ط	أَطَعْتُمْ
/IH/	Kasra ـِ	بِنت	/DH2/	(Thaa) ظ	الأَنظِمَة
/IY/	ـِي	فِيل	/AI/	(Ain) ع	الأَنْعَامَ
/IX/	ـِ	صِنف	/GH/	(Ghain) غ	الإِشغَال
/AW/	وَ	لَوم	/F/	(Faa) ف	الإِضافِيَّة
/AY/	يَ	ضَيف	/Q/	(Kaf) ق	الإِغلاق
/E/	(Hamza) ء	أَنشِطَةِ	/K/	(Kaf) ك	الإِلكتُرُونِيَّة
/B/	(Baa) ب	أَبدَت	/L/	(Lam) ل	آل
/T/	(Taa) ت	أَبرَمَت	/M/	(Meem) م	آمَالُ
/TH/	(Thaa) ث	أَثبَتَت	/N/	(Noon) ن	آمَنَ
/JH/	(Jeem) ج	أَجهِزَةُ	/H/	(Haa) هـ	آيَاتُهُ
/HH/	(Haa) ح	أَحجَام	/W/	(Waw) و	أَموَالَ
/KH/	(Khah) خ	أَخبَار	/Y/	(Yaa) ي	أَبُوظَبي

However, the Arabic phoneme group is considered to be good enough while we believe that this is still far from optimal group, and further efforts are needed to arrive at an optimized group of Arabic phonemes.

5.2.5 Arabic pronunciation dictionary

One of the main important components in ASRs is pronunciation dictionary (lexicon). It contains the phonetic transcriptions for all the words in the target language of the speech. A phonetic transcription is a sequence of phonemes that describes how the vocabulary should be pronounced. Ali et al. in [105] developed a software tool to generate pronunciation dictionaries for Arabic texts using Arabic pronunciation rules. AbuZeina et al. in [24, 53] developed this tool to take care of some cases such as within-word variation, and cross-word. We utilized this tool to generate the dictionary. The vocabulary list in the baseline dictionary is 14,234 words with within-word variants [53]. A sample from the pronunciation dictionary is listed in **Figure 5-1**.

أربَعَة	E AE R B AE AI AE H
أربَعَة(2)	E AE R B AE AI AE T
أربَعَةً	E AE R B AE AI AE T AE N
أربَعَةٍ	E AE R B AE AI AE T IH N

Figure 5-1 Sample from the pronunciation dictionary file

5.2.6 Arabic language model

Language model is another important component for ASR system. The CMU language toolkit (Open Source Toolkit for Speech Recognition) was used for language model training [38]. It is used to build a statistical language model from the transcription of the full diacritized transcription of the BCN and HQ speech. The way of building the language model consists of computation of the total count of 1-grams, 2-grams, and 3-grams from transcription text based and finally converting the n-grams to binary format of language model and standard ARPA format. For more information of language models, go back to chapter 2.

5.3 Experiments of Acoustic Modeling

In this section, we describe the experimental accuracy of the acoustic modeling proposed in chapter 4 for Arabic speakers. We also verify our proposed interpolation approaches for acoustic modeling by employing cross-validation test. All experiments were carried out using context dependent acoustic models with 8 Gaussians mixture.

In addition, the pronunciation model is rebuilt from transcription of broadcast news and the Qur'an corpus. So, the new lexicon has 107,914 words, and 28,845 vocabulary. Also, we used this lexicon within all experiments.

5.3.1 Experiment 1: Interpolation and Merging Approaches

The acoustic model interpolation and the merging approaches are tested by using the target (Arabic broadcast news) and source (the Holy Qur'an) phoneme matching which are the same in all experiments. Acoustic interpolation and merging were performed by using Arabic and the corresponding context dependent L1 acoustic model with 8 Gaussian mixtures. For acoustic model interpolation, Euclidean distance is used for measuring the distance between the Gaussians, and for acoustic model merging, the merging variant was applied. **Figure 5-2** and **Figure 5-3** show the WER of Arabic speakers by employing acoustic model interpolation and merging across varied weights.

In general, the results show that acoustic model merging performs better than acoustic model interpolation, although it creates a model with twice the number of Gaussian mixtures compared to the acoustic model interpolation. We note that when broadcast news weight is equal to 1.0, it is the baseline result. When broadcast news weight equals to 0.0, the broadcast news acoustic model is replaced by the corresponding phonemes from the Holy Qur'an (L1) acoustic model of the Arabic speaker.

The results indicate that replacing state of phoneme in target acoustic model with phoneme in source acoustic model has a good impact on the accuracy of speech recognition because of merging new state within target acoustic model.

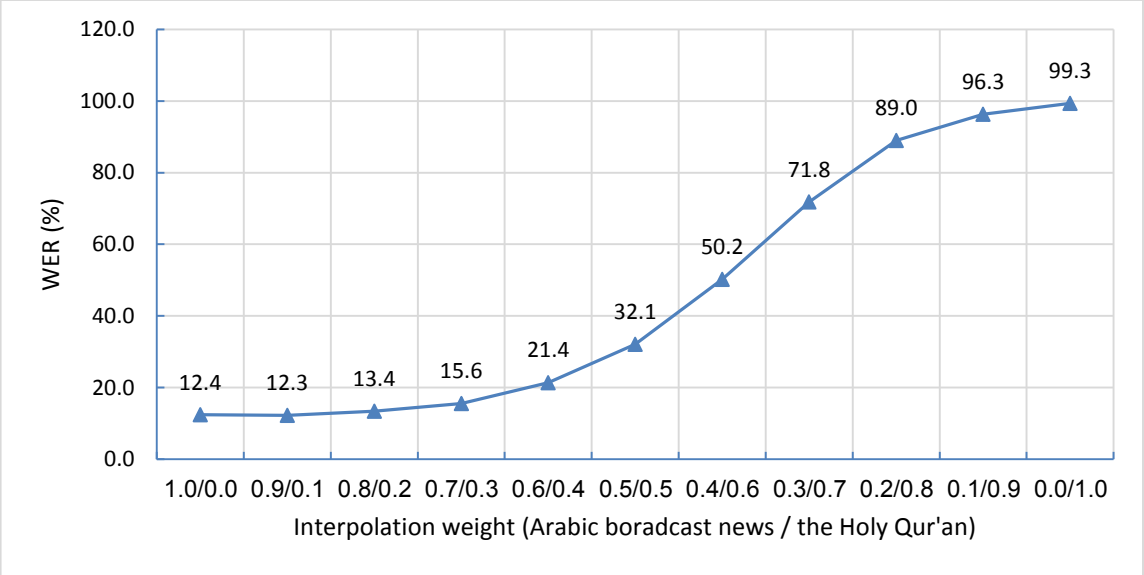

Figure 5-2: WER on Arabic speakers by interpolating acoustic models, which are created from a 8 Gaussian CD target (Arabic broadcast news) and source (the Holy Qur'an) acoustic models across different weights

Figure 5-2 shows the WER of Arabic speakers by employing interpolation of acoustic models across varied weights with broadcast news and the Holy Qur'an datasets. The figure shows that the proposed interpolated model has average WER which is always higher than the baseline. The interpolated model has the best average WER when the weights for broadcast news and the Holy Qur'an are at 0.9 and 0.1 respectively. The WER were 12.3%. On the other hand, the worst WER is when the weights for broadcast news and the Holy Qur'an are at 0.0 and 1.0. The WER were 99.3%.

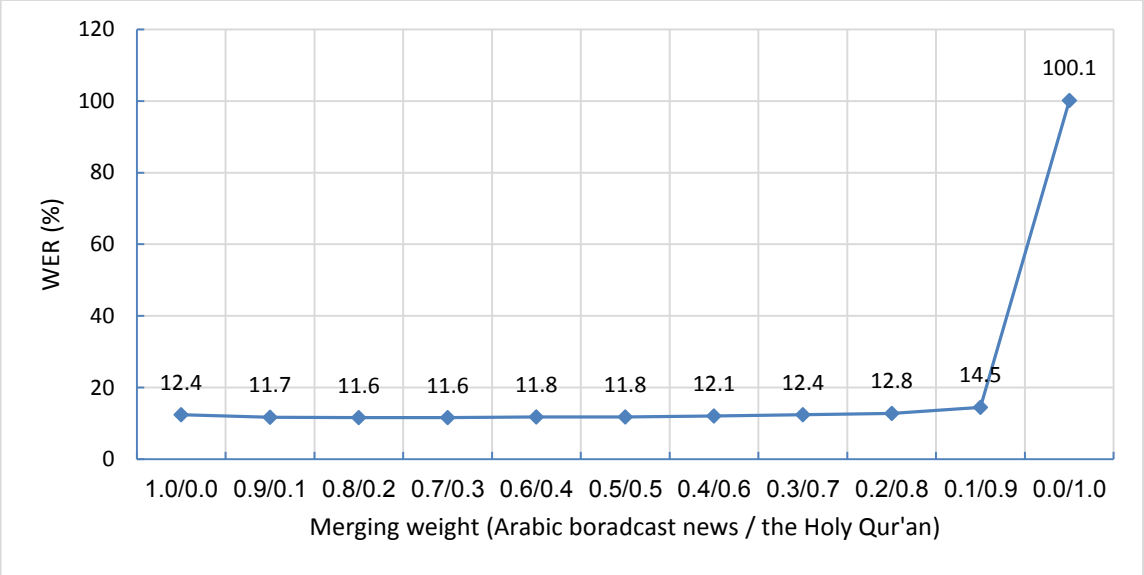

Figure 5-3: WER on Arabic speakers by merging acoustic models, which are created from a 8 Gaussian CD target (Arabic broadcast news) and source (the Holy Qur'an) acoustic models across different weights

Figure 5-3 shows that the merging model has the best average WER when the weights for broadcast news and the Holy Qur'an are at 0.8 and 0.2, 0.7 and 0.3 respectively. The WER were 11.6%. On the other hand, the worst WER is when the weights for broadcast news and the Holy Qur'an are at 0.0 and 1.0. The WER were 100.1%.

5.3.2 Experiment 2: Hybrid of Interpolation and Merging Approach for offline adaptation

In this section, the hybrid of interpolation and merging approaches proposed as seen in chapter 4 are applied to offline adapt the target acoustic model (Arabic broadcast news). Recall that the approach interpolates source and target Gaussian in the matching states that are near to each other. For Gaussians that are far from each other, the source or target Gaussians will be merged.

Euclidean distance was used as the distance measure. The initial context dependent acoustic models for Arabic broadcast news and the Holy Qur'an were employed for the experiments. The resulting models have an average of 16 Gaussians per state. Indeed, Sphinx-3 speech recognition system is not capable of handling varied number of

Gaussians per state. To model this, we set all states to the maximum number of Gaussian mixtures possible. Therefore, the means, variances and mixture weights for the empty Gaussians are set to zero.

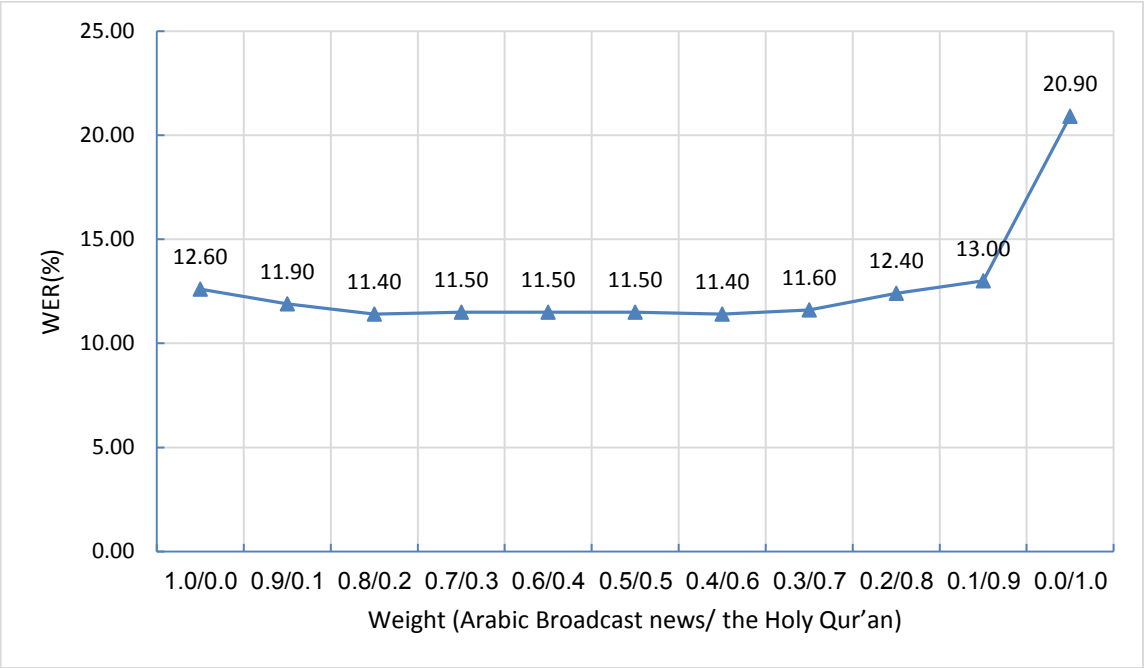

Figure 5-4 WER on Arabic speakers using hybrid model created from a 8 Gaussian CD Arabic broadcast news and the Holy Qur'an acoustic models with varied weights

The results from the experiments of adapting the target (Arabic broadcast news) acoustic model with L1 (the Holy Qur'an) acoustic model are very encouraging. In general, the results illustrate that using L1 for adaptation with the hybrid approach performs better than acoustic model merging in **Figure 5-3**.

Figure 5-4 shows that the proposed hybrid model has the lowest WER when the weights for broadcast news and the Holy Qur'an are at 0.8 and 0.2 respectively. The WER was 11.40%. On the other hand, the highest WER is when the weights for broadcast news and the Holy Qur'an are at 0.0 and 1.0. The WER was 20.90%.

It is also noteworthy that it produces an average relative WER improvement of 1% for Arabic speakers, while the acoustic merging approach has an average of 0.8% of improvement for Arabic speakers. In addition, the hybrid approach uses a smaller

number of Gaussians compared with merging approach. The results also indicates that replacing state of phoneme in target AM with phoneme in source AM has good impact on the accuracy of speech recognition because of integrating new state within target AM.

To sum up, the results show that the modeling are very promising with L1 acoustic model for adaptation of the target language. However, a small shortcoming is that the resulting model will have a higher number of Gaussian mixtures than the original target model.

5.3.3 Experiment 3: Validation of the Hybrid Approach

In this section, we went to verify the accuracy in experiment 2 (Hybrid of Acoustic Model Interpolation and Merging Approach) by using the cross-validation evaluation. Also, it is difficult to compare accuracy of WER produced and requires a cross-validation method to avoid different WER outcomes because of different initial partitions of dataset. To make certain results of the WER, 5-fold cross validation can be used to find a generalized, and optimal of the WER.

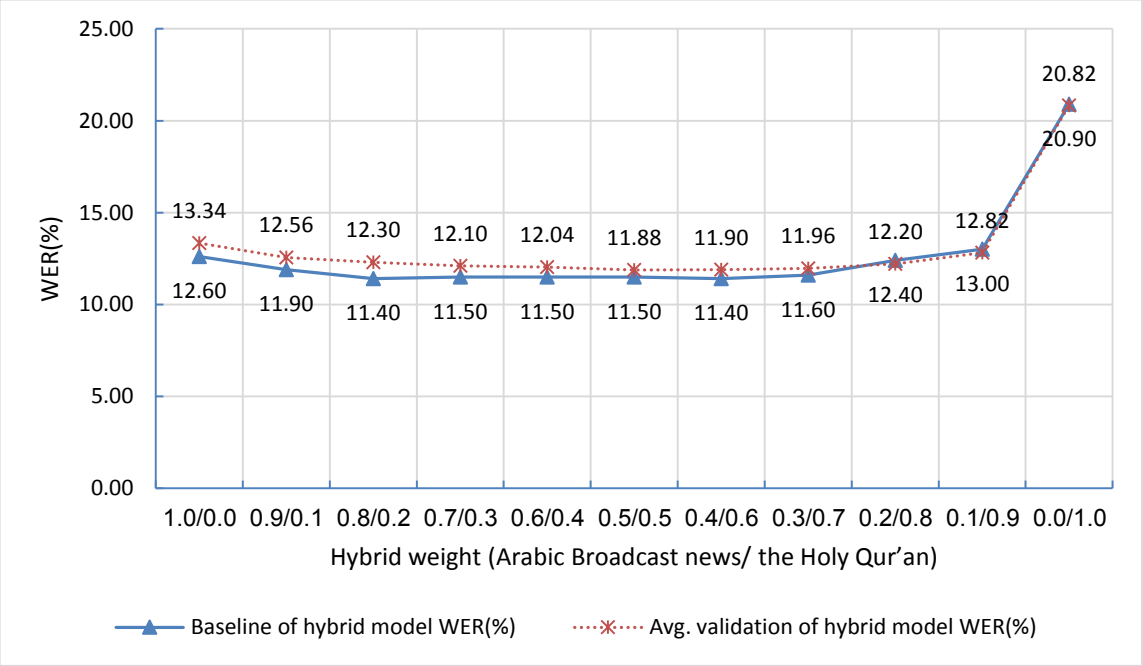

Figure 5-5 Validation of WER on the Hybrid Approach by Cross-Validation

Cross-validation attempts to mimic test data as follows: first, the sample is randomly split into 5 equal-sized, non-overlapping subsamples (5 is a frequent default). Second, each of the 5 subsamples is for training while a testing from the other m - 1 subsamples. Third, the predictive accuracy of each WER is tested with the one subsample not included in its training set. Finally, averaging over the m fit assessments. This proposed method can be used to determine a reasonable value for the WER.

Table 5-5 Validation of WER on the Hybrid Approach by Cross-Validation

Hybrid weight	Avg. Base WER (%)	Avg. WER (%)	Improvement (%)
1.0/0.0	13.28	13.34	-0.06
0.9/0.1	13.28	12.56	+0.72
0.8/0.2	13.28	12.3	+0.98
0.7/0.3	13.28	12.1	+1.18
0.6/0.4	13.28	12.04	+1.24
0.5/0.5	13.28	11.88	+1.4
0.4/0.6	13.28	11.9	+1.38
0.3/0.7	13.28	11.96	+1.32
0.2/0.8	13.28	12.2	+1.08
0.1/0.9	13.28	12.82	+0.46
0.0/1.0	13.28	20.82	-7.54

The results from the experiments of validation of Hybrid approach as seen in **Figure 5-5** and **Table 5-5** by cross-validation are very close to the results from of Hybrid approach. **Table 5-5** and **Figure 5-5** show that the hybrid model by cross-validation has the lowest average of WER when the weights for broadcast news and the Holy Qur'an are at 0.5 and 0.5 respectively. The WER was 11.88%. On the other hand, the highest WER is when the weights for broadcast news and the Holy Qur'an are at 0.0 and 1.0. The WER was 20.82%.

Overall, the average of results show that using L1 for adaptation with hybrid approach at 50%, performs the better result. It is also noteworthy that it produces an average significant WER improvement of 1.4% for Arabic speakers.

5.3.4 Experiment 4: Several Distance Equations within Hybrid Approach

In this section, we applied different equations to estimate the distance between sources and target Gaussian.

Table 5-6 and **Figure 5-6** show improving of the WER of Arabic speakers by employing hybrid approach across varied distance equations.

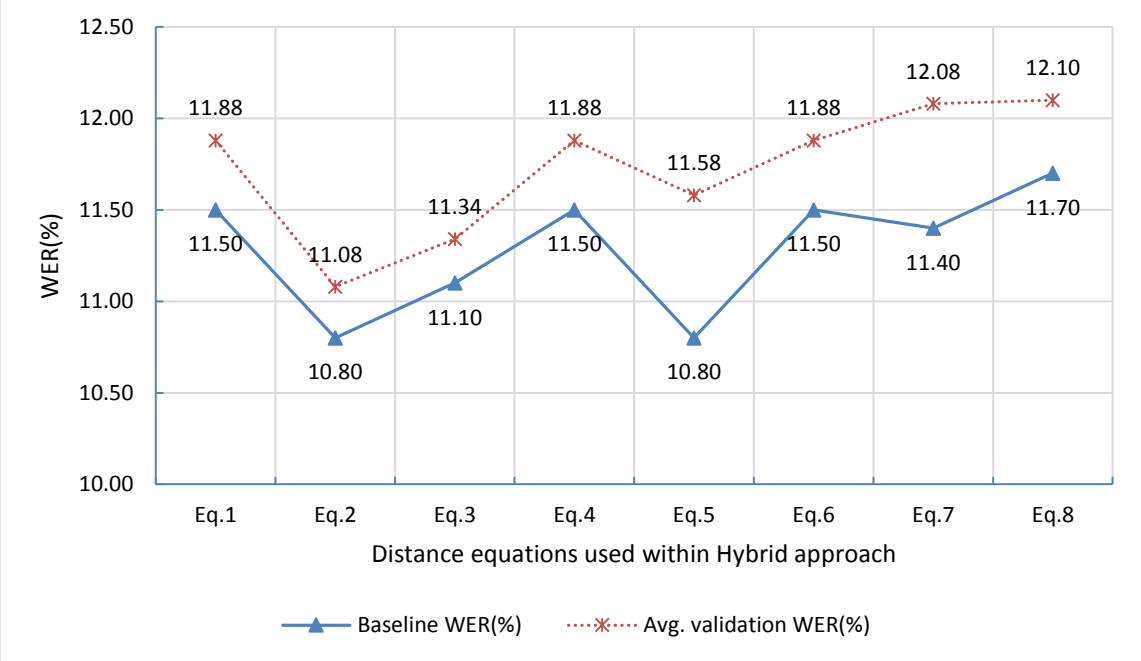

Figure 5-6: Several Distance Equations Used within Hybrid Approach. This Experiment was applied at Weight 0.5/0.5 which is the best result in experiment 3.

The proposed distance equation of Manhattan distance within Hybrid approach can produce very good recognition results if a suitable weight is assigned for modeling. However, in our experiment, we evaluated different weights as a priori.

Overall, the average of results show that the Manhattan distance is very promising within Hybrid approach to use as distance measure in next experiments. In most cases, Manhattan distance scores an average relative WER improvement of 2.2% for Arabic speakers in

Table 5-6 and **Figure 5-6**, while the other equations recorded less than Manhattan distance equation. Also, the best result at weight 0.5/0.5 in experiment 3 was used in this experiment for evaluating and selecting the best distance equation within Hybrid approach.

Table 5-6: Summary of Several Distance Equations Used within Hybrid Approach

Eq.#	Distance equation	Avg. Base WER (%)	Avg. WER (%)	Improve
1.	$d(p,q) = \sqrt{\sum (p-q)^2}$	13.28	11.88	+1.4
2.	$d(p,q) = \sum \|\|p-q\|\|$	13.28	11.08	+2.2
3.	$d^2(p,q) = \sum (p-q)^2$	13.28	11.34	+1.94
4.	$d(p,q) = \sqrt{\sum p^2 + q^2 - 2pq}$	13.28	11.88	+1.4
5.	$d(p,q) = \sqrt{\sum (p^2 + q^2)}$	13.28	11.58	+1.7
6.	$d(p,q) = \sqrt[2]{\sum (p-q)^2}$	13.28	11.88	+1.4
7.	$d(p,q) = \sqrt[3]{\sum (p-q)^3}$	13.28	12.08	+1.2
8.	$d(p,q) = \sqrt[4]{\sum (p-q)^4}$	13.28	12.10	+1.18

5.3.5 Experiment 5: Manhattan distance within Hybrid Approach for offline adaptation

In this experiment, Manhattan distance is applied within the hybrid of interpolation and merging approach. Manhattan distance was used as the distance measure. The initial context dependent acoustic models for Arabic broadcast news and the Holy Qur'an were employed for the experiments.

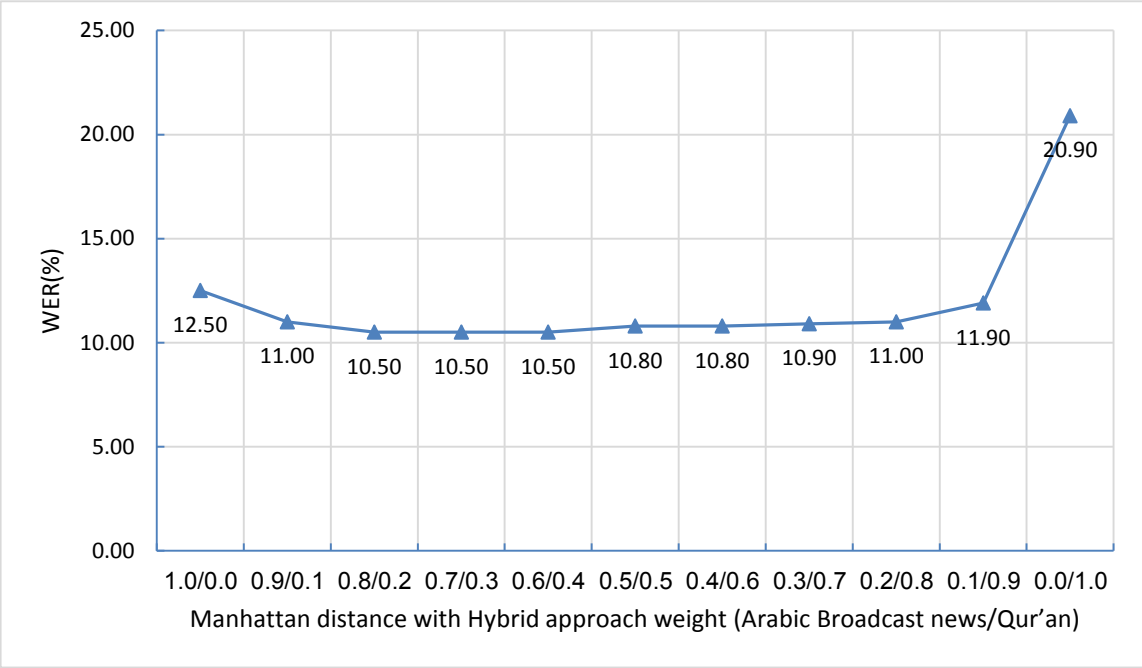

Figure 5-7 WER on Arabic speakers using Manhattan distance within hybrid models created from a 8 Gaussian CD Arabic broadcast news and Holy Qur'an acoustic models with varied weights

The results from the experiments show that Manhattan distance within Hybrid approach performs better than Euclidean distance within Hybrid approach. **Figure 5-7** shows that the proposed model has the lowest average WER when the weights for broadcast news and the Holy Qur'an are at 0.8/0.2, 0.7/0.3, and 0.6/0.4 respectively. The WER was 10.50%. On the other hand, the highest WER is when the weights for broadcast news and the Holy Qur'an are at 0.0 and 1.0. The WER was 20.90%.

Overall, the key point to note is that WER is significantly reduced by 1.9% when the Arabic broadcast news weight equals 0.8, 0.7, and 0.6.

5.3.6 Experiment 6: Validation of the Manhattan distance within Hybrid Approach

In this experiment, we validated the result from experiment 5 (Manhattan distance within Hybrid Approach) on 5 cross-validation corpus by conducting the test on Arabic speakers. 5-cross corpus was used to create CD models with 1138 states. To verify that these results are not just due to some peculiarities of the Arabic speech, the same experiments are redone with different training and testing datasets of speech in **Figure 5-8**.

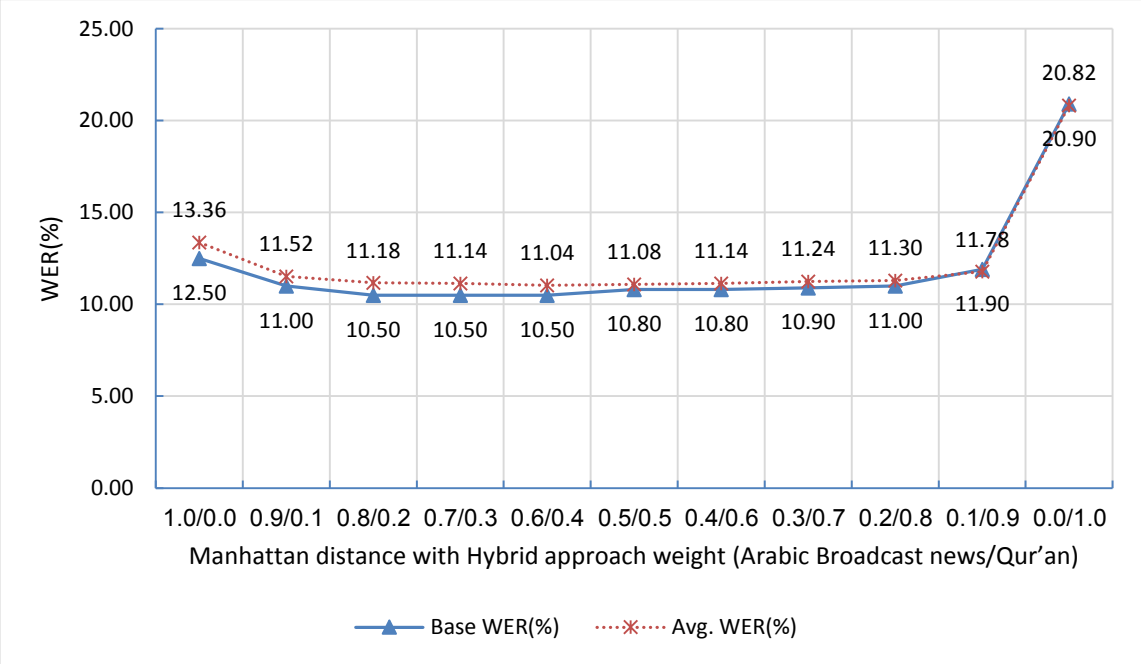

Figure 5-8 Validation of the WER on Arabic speakers using Manhattan distance with varied weights using cross-validation

Figure 5-8 compares the average WERs of an ASR system employing the Manhattan distance within hybrid approach with 5-cross dataset and the Manhattan distance within hybrid approach with broadcast news dataset. **Figure 5-8** shows that Manhattan distance within hybrid approach provided a relative average WER reduction of 2.24% for Arabic speech when compared to the average baseline ASR system.

Figure 5-8 and **Table 5-7** show that the best WER for Arabic broadcast news speakers was achieved when CD weight is 0.6. The WER is 11.04%.

Consequently, the final acoustic models which were obtained encompassed both the characteristics of broadcast news speech and the Qur'an speech. The results from the experiments show that an ASR system employing the proposed Manhattan distance within Hybrid method relatively reduced the average WER by 2.24%, when compared to Euclidian distance within Hybrid method.

Table 5-7 Validation of the WER on Arabic speakers using Manhattan distance within hybrid models created from a 8 Gaussian CD Arabic broadcast news and Holy Qur'an acoustic models with varied weights using cross-validation

Hybrid weight	Avg. Base WER (%)	Avg. WER (%)	Improve (%)
1.0/0.0	13.28	13.36	-0.08
0.9/0.1	13.28	11.52	+1.76
0.8/0.2	13.28	11.18	+2.1
0.7/0.3	13.28	11.14	+2.14
0.6/0.4	13.28	11.04	+2.24
0.5/0.5	13.28	11.08	+2.2
0.4/0.6	13.28	11.14	+2.14
0.3/0.7	13.28	11.24	+2.04
0.2/0.8	13.28	11.3	+1.98
0.1/0.9	13.28	11.78	+1.5
0.0/1.0	13.28	20.82	-7.54

5.3.7 Conclusions from Acoustic Modeling

We have examined the proposed Arabic acoustic modeling approaches to adapt Arabic acoustic model for Arabic speakers without using any Arabic speech from the same domain such as broadcast news in our case. One type of speech was experimented for adaptation. For Arabic speakers, they were Arabic (L1). Interestingly, with

appropriate weight to the source model, formal Arabic speech appears to be useful for adapting another formal Arabic speaker in another domain.

The hybrid approach performs better than the acoustic model interpolation and merging approaches given the source L1 acoustic model. It has also shown to be beneficial with Manhattan distance.

Moreover, the hybrid approach proposed for Arabic acoustic modeling has proven to be also useful for modeling between two different contexts to achieve an intermediate state, where the resulting model reduces the errors of Arabic speakers. Also, this result supports many previous efforts which indicated that the Hybrid of acoustic modeling interpolation and merging have good impact on improving speech recognition when it is used between two acoustic models in the same language. The results also show that the hybrid approach is very promising with L1 acoustic model for adaptation target language. However, a small shortcoming is that the resulting model will have a higher number of Gaussian mixtures than the original target model.

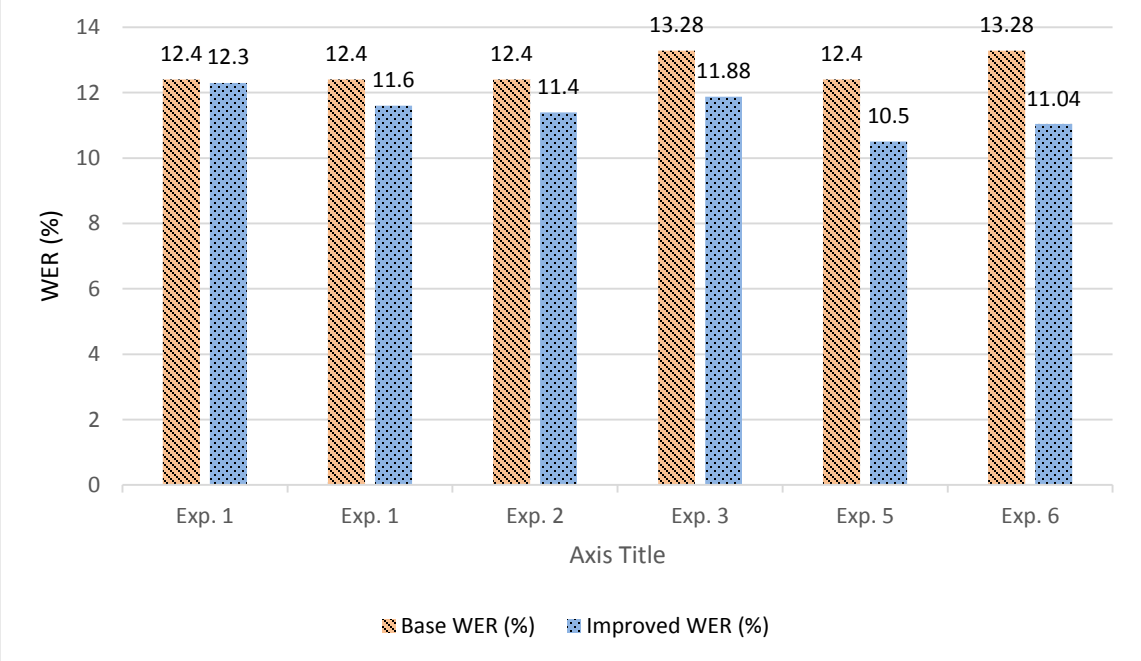

Figure 5-9 Summary of all experiments containing the best WER of acoustic model adaptation methods

In addition, this research proposed a new method for evaluating our results which is cross-validation evaluation. This new proposed method can show a reasonable value for the WER.

Finally, **Table 5-8** and **Figure 5-9** show summary all experiments containing the best WER of acoustic model adaptation methods.

Table 5-8 Summary of all experiments containing the best WER of acoustic model adaptation methods.

Exp. No.	Experiment name	Base WER (%)	Best Avg. WER (%)	Improve (%)
Exp. 1	Interpolation method.	12.4	12.3	+0.1
Exp. 1	Merging method.	12.4	11.6	+0.8
Exp. 2	Hybrid of interpolation and merging method.	12.4	11.4	+1
Exp. 3	Hybrid of interpolation and merging method with cross-validation dataset.	13.28	11.88	+1.4
Exp. 5	Manhattan distance with hybrid method.	12.4	10.5	+1.9
Exp. 6	Manhattan distance with hybrid method with cross-validation dataset.	13.28	11.04	+2.24

5.4 Experiments of Language Modeling

In this section, we describe the experiments of performance of language modeling proposed for Arabic language. For an optimal adaptation of language models in specific domains it is required that the system has a previous knowledge of data belonging to the same domain or, at least, to related one. Precisely, the aim in statistical language model adaptation is to add new sources of information to the previously existent models with the objective of enriching them. Echeverry-Correa et al. in [92] show that the goal in LM adaptation is to reflect the changes that the language experiences when moving towards different domains or, as in some applications, when dealing with multiple speakers.

5.4.1 Experiment 7: Interpolation Language Model Approach

In this experiment, the interpolation approach is applied between the broadcast news transcription and the Holy Qur'an transcription. **Figure 5-10** presents different interpolation values between two language models which are Arabic broadcast news and the Holy Qur'an.

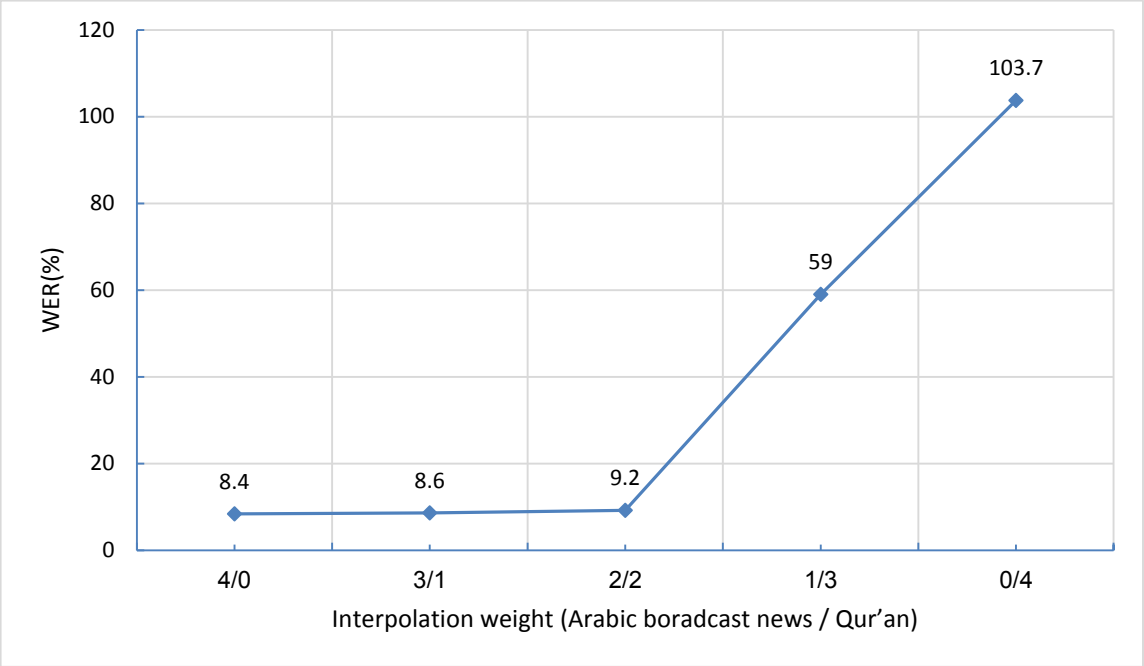

Figure 5-10 WER on Arabic speakers using interpolation language models, which are Arabic broadcast news and Qur'an language models

Our results show that the expanded language model with HQ transcription does not add appreciable improvements. However, replicating BCN transcription four times performs better than replicating BCN transcription three times with HQ transcription one time. Overall, **Figure 5-10** shows that, the best WER for Arabic speakers achieved the interpolation weights for broadcast news and the Holy Qur'an are at 4 and 0 respectively. The WER was 8.4%.

5.4.2 Experiment 8: Interpolation Language Model Approach followed by Manhattan distance within Hybrid Approach

In this experiment, the interpolation language model approach and Manhattan distance within Hybrid approach are applied together to evaluate integrating two approaches on the WER of Arabic ASR system.

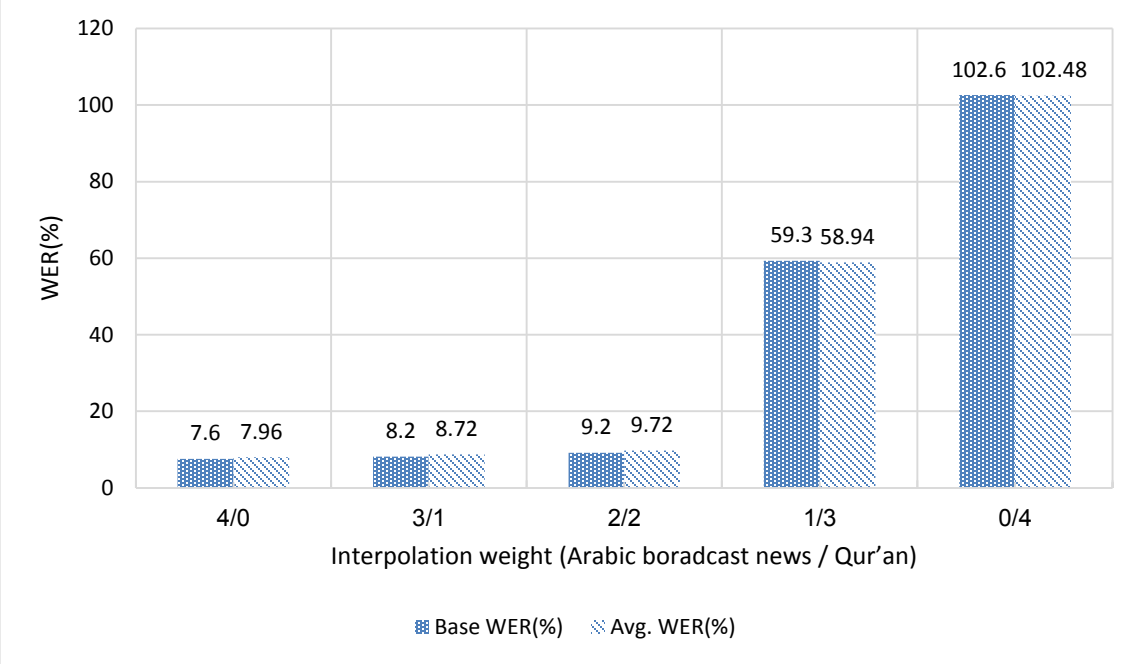

Figure 5-11 WER on Arabic speakers using interpolation language models

Overall, **Figure 5-11** shows that the average word error rate is significantly reduced by 4.8% when the interpolation language model and Hybrid of Interpolation and Merging Approach with Manhattan distance is applied. Note that when Arabic broadcast news weight is equal to 4 performs better than other weights. The best WER for Arabic speakers was achieved when interpolation weight was 4 for broadcast news. The WER is 7.6%.

5.4.3 Conclusions from Language Modeling

We have examined interpolation of language modeling. Adding new sources of information to the previously existent models with the objective of enriching them reduces the WER more than the base of the WER. Interestingly, replicating BCN

transcription four time shows that this method can significantly reduce the WER for Arabic speakers. However, the HQ transcription does not add appreciable improvements when it is added to language model.

Our result show that the best accuracy of WER is 7.6% when we applied hybrid of acoustic model, and interpolation of language model together. The proposed method is significantly reduced WER by 4.8%, 5.32% based on baseline, and cross-validation dataset respectively.

5.5 Experiments of Pronunciation Modeling

This section describes the rules used in this work for generating new variants in lexicon. A rule based approach is used to automatically generate lexicon dictionary for a given diacritized Arabic transcription.

5.5.1 Experiment 9: removing all diacritized text

The pronunciation model was tested by removing all diacritized text from pronunciation dictionary to work without diacritized text. The result of the experiment shows increased in the WER to reach 90%.

5.5.2 Experiment 10: eliminating all duplicate in pronunciation the word

The pronunciation model was tested by eliminating all duplicate in pronunciation of the word from pronunciation dictionary. The result of the experiment shows an increase in the WER to reach 30%.

5.5.3 Experiment 11: add Al-Shamsi and Al-Moon

The pronunciation model was tested by adding Al-Shamsi (ال الشمسية) and Al-Moon (ال القمرية) in pronounce the word in pronunciation dictionary. The result of the experiment shows an increase in the WER to reach 28.9%.

5.5.4 Experiment 12: replacing FATHA followed by WAW to WAW

The pronunciation model was tested by replacing FATHA followed by WAW (/AW/) to WAW (/W/) in pronunciation of the word in pronunciation dictionary. The result of the experiment shows an increase in the WER to reach 14.5%.

5.5.5 Experiment 13: splitting WAW rule

The pronunciation model was tested by adding two rules to pronounce the WAW (/W/). Firstly, DAMMA followed by Vowels (FATHA, DAMMA, KASRA, and SHADDA) will generate WAW (/W/). Secondly, FATHA followed by Vowels (FATHA, DAMMA, KASRA, and SHADDA) will generate FATHA followed by WAW (/AW/). The result of the experiment shows an increase in the WER to reach 14.1%.

5.5.6 Experiment 14: Unifying the pronunciation of Tanween

The pronunciation model was tested by unified pronounce of Tanween. The result of the experiment shows an increase in the WER to reach 14.1%.

5.5.7 Experiment 15: merging the pronunciation of FATHA, Long FATHA, Pharyngeal Version of FATHA, and Long Version of Pharyngeal Version of FATHA

The pronunciation model was tested by merging the pronunciations of Emphatic Version of FATHA (/AH/), Long FATHA (/AH:/), Pharyngeal Version of FATHA (/AA/), and Long Version of Pharyngeal Version of FATHA (/AA:/) to become FATHA (/AE/). These rules updated in FATHA and FATHATAN rules. The result of the experiment shows an increase in the WER to reach 14.2%.

5.5.8 Experiment 16: converting the pronunciation of the Pharyngeal Version of DAMMA to DAMMA

The pronunciation model was tested by converting the pronunciation of the Pharyngeal Version of DAMMA (UX/) to DAMMA (UH). This rule updated in DAMMA, and DAMMATAN rules. In addition, we deleted Pharyngeal Version of DAMMA (UX/) from dictionary. The result of the experiment shows an increase in the WER to reach 14.3%.

5.5.9 Experiment 17: converting the pronunciation of Pharyngeal Version of KASRA to KASRA

The pronunciation model was tested by converting the pronunciation of Pharyngeal Version of KASRA (/IX/) to KASRA (/IH/). This rule updated in KASRATAN rules. In

addition, we deleted Pharyngeal Version of KASRA (/IX/) from dictionary. The result of the experiment shows an increase in the WER to reach 14.3%.

The **Table 5-9** shows that the proposed pronunciation rules have average WER which is mostly higher than the baseline.

Table 5-9 WER on Arabic speakers using pronunciation rules in pronunciation modeling

Experiment #	WER (%)	Improve
9	88.0	-75.6
10	28.0	-15.6
11	26.9	-14.5
12	12.5	-0.1
13	12.1	+0.3
14	12.1	+0.3
15	12.2	+0.2
16	12.3	+0.1
17	12.3	+0.1
18	12.3	+0.1

5.5.10 Conclusions from Pronunciation Modeling

We have examined different rules in pronunciation modeling. Adding variants created from new rules into pronunciation dictionary increase the WER more than the base of the WER. That means our proposed phonetic rules cannot describe variations in speech accurately. In other words, we need domains' expertise in the Arabic language to generate phonetic rules

In our work pronunciation rules are manually derived for the modification of lexicons. However, this approach requires language expertise to reflect the rules of pronunciation accurately which are called to find the possible variants. However, this approach depends on the language itself to provide the phonological rules.

5.6 Comparison with Other approaches

Based on our experiments that were carried out using broadcasting news, we decided to compare our results based on the BCN corpus with the available related work models that present their experiment results based on the same corpus. As shown in **Table 5-10**, we achieved a lower rate of WER.

Table 5-10 Comparison with other models

Authors	Base WER (%)	WER (%)	Improve (%)	Approach
Our model	13.28	7.96	+5.32	Hybrid of interpolation and merging of acoustic model followed by Interpolation of language model.
AbuZeina et al. [53]	13.39	11.17	+2.22	A direct data-driven.
AbuZeina et al. [24]	12.21	9.91	+2.3	A knowledge-based.
AbuZeina et al. [137]	13.39	11.23	+2.16	A data-driven.
AbuZeina et al. [138]	12.21	9.82	+2.39	Part of speech tagging.

CHAPTER 6

Conclusions and Future Works

Conclusions

Automatic speech recognition applications have been implemented in various domains. However, ASR still suffers from different difficulties in treating Arabic speech. The misrecognition is due to the variation in Arabic speech which exists at phonetic, word, or sentence level. This type of error can occur in pronunciation, acoustic, and language models.

The main existing approaches to acoustic modeling can be classified to acoustic model reconstruction, acoustic model interpolation, acoustic model merging, hybrid interpolation and merging of acoustic model and the general speaker adaptation approaches. These approaches either use a small amount of speech or another speech for adaptation because of the difficulty to acquire a large amount of speech. We have proposed to use Arabic resource (L1) for adapting target acoustic model for Arabic speakers. Three types of resources have been identified. They are the native language of the speaker (L1), any non-native language (L2), and languages close to the native language of the speakers (L3). Besides, existing approaches do not address the Arabic language for acoustic modeling.

For studying Arabic speech recognition, we have collected an Arabic speech corpus from the Holy Qur'an. The corpus was uttered by one Arabic speaker. So, we have used this resource for proposing different acoustic modeling approaches. If several acoustic models are available, the hybrid of acoustic model interpolation and merging has proven to be useful for modeling Arabic acoustic. The idea is to interpolate the target and source Gaussian that are close, and merge them if they are far from each other. We have also shown that the hybrid approach is better than the interpolation or merging approaches. The results show that, by giving appropriate (or intermediate) weight, the approach reduced the error rate of Arabic speech recognition system. Overall, the HQ as L1

resources that have been proposed, can perform a good result for adaptation. On the other hand, our improved hybrid approach is better than the hybrid approach.

Besides, in cases where several corpora are available, we have suggested interpolation approach for adaptation language model. It is easier to carry out. The results show that, by giving appropriate (or intermediate) weight, the approach reduced the error rate of Arabic speech recognition system.

In this study, pronunciation modeling is generally based on data-driven (manually) approach. This method, nevertheless, did not lead to a significant improvement. Our experiments showed that the data-driven approach may fail to reduce the error rate of Arabic speech recognition system when language expertise and Arabic speech are limited.

Future works

1. In future work, we suggest to address adapting the acoustic model using other resources such as any non-native language spoken by the same Arabic speakers (L2) and language close to the Arabic language of the speaker (L3).
2. The hybrid method of interpolation and merging is a promising approach for modeling Arabic acoustic models. We will suggest an automatic way to estimate the weights given to some Arabic speech.
3. Further, we will propose to use Arabic transcription without diacritical marks for adapting pronunciation, acoustic and language models in which we can benefit from huge Arabic transcription without diacritical marks. Moreover, we will propose more accurately phonological rules to improve pronunciation model.
4. We recommend sharing Arabic language experts to help in the construction of the pronunciation rules to be accurate especially in pronunciation model.
5. In addition, we suggest to collect more than one speaker in the Holy Qur'an speech dataset. Furthermore, we recommend to use other intelligent methods such as deep learning.

Bibliography

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85-100, 1// 2014.

[2] C. Cai, Y. Xu, D. Ke, and K. Su, "A Fast Learning Method for Multilayer Perceptrons in Automatic Speech Recognition Systems," *Journal of Robotics*, vol. 2015, 2015.

[3] M. Khasawneh, K. Assaleh, W. Sweidan, and M. Haddad, "The application of polynomial discriminant function classifiers to isolated Arabic speech recognition," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, 2004, pp. 3077-3081 vol.4.

[4] A. Gorin, "Acoustic Model Structuring for Improving Automatic Speech Recognition Performance," Doctoral dissertation, University of Lorraine, 2014.

[5] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-vocabulary dictation using SRI's DECIPHER speech recognition system: progressive search techniques," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 1993, pp. 319-322 vol.2.

[6] L. Lin-Shan, "Voice dictation of Mandarin Chinese," *Signal Processing Magazine, IEEE*, vol. 14, pp. 63-101, 1997.

[7] A. Lee, T. Kawahara, and K. Shikano, "Julius-an open source real-time large vocabulary recognition engine," in *EUROSPEECH2001*, Aalborg, Denmark, 2001, pp. 1691-1694.

[8] S. Furui, "Automatic speech recognition and its application to information extraction," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 11-20.

[9] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, *et al.*, "JUPlTER: a telephone-based conversational interface for weather information," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, pp. 85-96, 2000.

[10] A. Gruenstein, J. Orszulak, S. Liu, S. Roberts, J. Zabel, B. Reimer, *et al.*, "City browser: developing a conversational automotive HMI," presented at the CHI '09 Extended Abstracts on Human Factors in Computing Systems, Boston, MA, USA, 2009.

[11] A. Franz and B. Milch, "Searching the Web by voice," presented at the Proceedings of the 19th international conference on Computational linguistics - Volume 2, Taipei, Taiwan, 2002.

[12] J. R. Glass, T. J. Hazen, D. S. Cyphers, K. Schutte, and A. Park, "The MIT spoken lecture processing project," presented at the Proceedings of HLT/EMNLP on Interactive Demonstrations, Vancouver, British Columbia, Canada, 2005.

[13] K. Harrenstien. (2009, 1, May). *Automatic captions in YouTube*. Available: https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html

[14] F. Metze, C. Fugen, Y. Pan, and A. Waibel, "Automatically Transcribing Meetings using Distant Microphones," in *ICASSP*, 2005, pp. 989-992.

[15] M. A. Peabody, "Methods for pronunciation assessment in computer aided language learning," Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 2011.
[16] S. M. Witt and S. J. Young, "Language learning based on non-native speech recognition," in *Eurospeech*, 1997.
[17] Y. Xu, A. Goldie, and S. Seneff, "Automatic question generation and answer judging: a q&a game for language learning," in *SIGSLaTE*, 2009, pp. 57-60.
[18] W. Wahlster, *Verbmobil: foundations of speech-to-speech translation*: Springer Science & Business Media, 2013.
[19] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, *et al.*, "Janus-III: speech-to-speech translation in multiple languages," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, pp. 99-102 vol.1.
[20] F. Hamidi, M. Baljko, N. Livingston, and L. Spalteholz, "CanSpeak: A Customizable Speech Interface for People with Dysarthric Speech," in *Computers Helping People with Special Needs*. vol. 6179, K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 605-612.
[21] D. Jones, E. Gibson, W. Shen, N. Granoien, M. Herzog, D. Reynolds, *et al.*, "Measuring human readability of machine generated text: three case studies in speech recognition and machine translation," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. v/1009-v/1012 Vol. 5.
[22] I. McGraw, C.-y. Lee, I. L. Hetherington, S. Seneff, and J. Glass, "Collecting Voices from the Cloud," in *LREC*, 2010.
[23] G. A. Sanders, A. N. Le, and J. S. Garofolo, "Effects of word error rate in the DARPA communicator data during 2000 and 2001," in *INTERSPEECH*, 2002, p. 2.
[24] D. AbuZeina, W. Al-Khatib, M. Elshafei, and H. Al-Muhtaseb, "Cross-word Arabic pronunciation variation modeling for speech recognition," *International Journal of Speech Technology*, vol. 14, pp. 227-236, 2011/09/01 2011.
[25] H. Hyassat and R. Abu Zitar, "Arabic speech recognition using SPHINX engine," *International Journal of Speech Technology*, vol. 9, pp. 133-150, 2006/12/01 2008.
[26] K. C. Ryding, *A reference grammar of modern standard Arabic*: Cambridge University Press, 2005.
[27] D. AbuZeina and M. Elshafei, *Cross-Word Modeling for Arabic Speech Recognition*: Springer Science & Business Media, 2011.
[28] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.
[29] S. Young, "A review of large-vocabulary continuous-speech recognition," *Signal Processing Magazine, IEEE*, vol. 13, pp. 45-47, 1996.
[30] N. Morgan and H. Bourlard, "Continuous speech recognition," *Signal Processing Magazine, IEEE*, vol. 12, pp. 24-42, 1995.
[31] F. Jelinek, *Statistical methods for speech recognition*: MIT press, Cambridge, MA, 1997.

[32] J. K. Baker, "Stochastic modeling for automatic speech understanding," in *Reddy R (ed) Speech recognition. Academic, New York*, ed, 1975, pp. 521–542.
[33] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.
[34] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition* vol. 14: PTR Prentice Hall Englewood Cliffs, 1993.
[35] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, *et al.*, *The HTK Book* vol. 3: Cambridge University Engineering Department, 2005.
[36] HTK. (2015, 4 Feb). Available: http://htk.eng.cam.ac.uk/
[37] L. KF, "Large vocabulary speaker independent continuous speech recognition: the Sphinx system," Doctoral dissertation, Carnegie Mellon University, 1988.
[38] CMU Sphinx. (2015, 1 Feb). Available: http://cmuSphinx.sourceforge.net/wiki/download
[39] P. Lamere, P. Kwok, W. Walker, E. B. Gouvêa, R. Singh, B. Raj, *et al.*, "Design of the CMU sphinx-4 decoder," in *Proceedings of the 8th European conference on speech communication and technology*, Geneva, Switzerland, 2003, pp. 1181–1184.
[40] X. D. Huang, "Phoneme classification using semicontinuous hidden Markov models," *Signal Processing, IEEE Transactions on*, vol. 40, pp. 1062-1067, 1992.
[41] G. D. Forney, Jr., "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, pp. 268-278, 1973.
[42] R. Mehla, Mamta, and R. K. Aggarwal, *Automatic Speech Recognition: A Survey* vol. 3, 2014.
[43] W. Ghai and N. Singh, "Literature review on automatic speech recognition," *International Journal of Computer Applications*, vol. 41, pp. 42-50, 2012.
[44] D. E. M. Abuzeina, "Utilizing data-driven and knowledge-based techniques to enhance arabic speech recognition," Doctoral dissertation, King Fahd University of Petroleum and Minerals (Saudi Arabia), 2011.
[45] T. Al Hanai, "Lexical and Language Modeling of Diacritics and Morphemes in Arabic Automatic Speech Recognition," Doctoral dissertation, Massachusetts Institute of Technology, 2014.
[46] T. Tien Ping, "Automatic Speech Recognition for Non-Native Speakers," Doctoral dissertation, Université Joseph-Fourier - Grenoble I, 2008.
[47] R. D. Kent, *Acoustic analysis of speech*: Canada, Singular Publishing Group 2002.
[48] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*: New Jersey, Prentice Hall PTR, 2001.
[49] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, pp. 357-366, 1980.
[50] Z. Tychtl and J. Psutka, "Speech production based on the mel-frequency cepstral coefficients," in *EuroSpeech*, 1999, pp. 2335-2338.

[51] M. A.-A. M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus," *Int. Arab J. Inf. Technol.*, vol. 9, pp. 84-93, 2012.

[52] K. Nahar, H. Al-Muhtaseb, W. Al-Khatib, M. Elshafei, and M. Alghamdi, "Arabic Phonemes Transcription using Data Driven Approach," *International Arab Journal of Information Technology (IAJIT)*, vol. 12, 2015.

[53] D. AbuZeina, W. Al-Khatib, M. Elshafei, and H. Al-Muhtaseb, "Within-word pronunciation variation modeling for Arabic ASRs: a direct data-driven approach," *International Journal of Speech Technology*, vol. 15, pp. 65-75, 2012/06/01 2012.

[54] T. Tien-Ping, L. Besacier, and B. Lecouteux, "Acoustic model merging using acoustic models from multilingual speakers for automatic speech recognition," in *Asian Language Processing (IALP), 2014 International Conference on*, 2014, pp. 42-45.

[55] N. Morales, J. H. Hansen, and D. T. Toledano, "MFCC Compensation for Improved Recognition of Filtered and Band-Limited Speech," in *ICASSP (1)*, 2005, pp. 521-524.

[56] Z. Heiga and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 3844-3848.

[57] G. Hinton, D. Li, Y. Dong, G. E. Dahl, A. Mohamed, N. Jaitly, *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *Signal Processing Magazine, IEEE*, vol. 29, pp. 82-97, 2012.

[58] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, pp. 328-339, 1989.

[59] J. Wu and C. Chan, "Isolated word recognition by neural network models with cross-correlation coefficients for speech dynamics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, pp. 1174-1185, 1993.

[60] T. R. Moore, "Twenty things we still don't know about speech proc. CRIM," in *FORWISS Workshop on Progress and Prospects of speech Research and Technology*, 1994.

[61] The CMU Pronunciation Dictionary. (2015, 1 October). Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[62] A. Senior, H. Sak, and I. Shafran, "Context dependent phone models for LSTM RNN acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4585-4589.

[63] M.-Y. Hwang, "Subphonetic acoustic modeling for speaker-independent continuous speech recognition," Doctoral dissertation, Computer science, School of Computer Science, Carnegie Mellon University, 1993.

[64] M. Alghamdi, M. Elshafei, and H. Al-Muhtaseb, "Arabic broadcast news transcription system," *International Journal of Speech Technology*, vol. 10, pp. 183-195, 2007/12/01 2009.

[65] X. Huang, A. Acero, and H.-W. Hon, "Spoken language processing," ed: Prentice Hall Englewood Cliffs, 2001.
[66] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-cambridge toolkit," in *Eurospeech*, 1997, pp. 2707-2710.
[67] D. AbuZeina, H. Al-Muhtaseb, and M. Elshafei, *Cross-Word Arabic Pronunciation Variation Modeling Using Part of Speech Tagging*, 2012.
[68] M. S. Seigel, "Confidence Estimation for Automatic Speech Recognition Hypotheses," Doctoral dissertation, Department of Engineering, University of Cambridge, 2013.
[69] L. Al-Sulaiti and E. S. Atwell, "The design of a corpus of Contemporary Arabic," *International Journal of Corpus Linguistics*, vol. 11, pp. 135-171, 2006.
[70] U. Uebler and M. Boros, "Recognition of non-native German speech with multilingual recognizers," in *Eurospeech*, 1999.
[71] L. M. Tomokiyo and A. Waibel, "Adaptation methods for non-native speech," *Multilingual Speech and Language Processing*, p. 6, 2001.
[72] S. Sam, E. Castelli, and L. Besacier, "Online unsupervised multilingual acoustic model adaptation for nonnative ASR," *ASEAN Engineering Journal*, vol. 1, pp. 76-86, 2012.
[73] Y. R. Oh, J. S. Yoon, and H. K. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition," *Speech Communication*, vol. 49, pp. 59-70, 1// 2007.
[74] Y. Liu and P. Fung, "Partial Change Accent Change Models for Accented Mandarin Speech Recognition, ASRU'03: St," *Thomas, US Virgin Islands*, pp. 111-113, 2003.
[75] Y. Liu and P. Fung, "Modeling partial pronunciation variations for spontaneous Mandarin speech recognition," *Computer Speech & Language*, vol. 17, pp. 357-379, 10// 2003.
[76] S. M. Witt and S. J. Young, "Off-line acoustic modelling of non-native accents," in *Eurospeech*, Budapest, Hungary, 1999, pp. 1367-1370.
[77] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Doctoral dissertation, University of Cambridge, 1999.
[78] Z. Wang and T. Schultz, "Non-native spontaneous speech recognition through polyphone decision tree specialization," in *INTERSPEECH*, 2003.
[79] S. Steidl, G. Stemmer, C. Hacker, and E. Nöth, "Adaptation in the pronunciation space for non-native speech recognition," in *Proc. ICSLP*, South Korea, 2004, pp. 2901-2904.
[80] T. Tien-Ping and L. Besacier, "Acoustic Model Interpolation for Non-Native Speech Recognition," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-1009-IV-1012.
[81] T. P. Tan and L. Besacier, "Modeling context and language variation for non-native speech recognition," in *Interspeech*, Antwerp, Belgium, 2007, pp. 1429-1432.
[82] J. J. Morgan, "Making a speech recognizer tolerate non-native speech through Gaussian mixture merging," in *InSTIL/ICALL Symposium 2004*, Venice, Italy, 2004.

[83] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints," in *The Ninth International Conference on Spoken Language Processing-ICSLP 2006*, Pittsburgh, 2006, pp. 109-112.

[84] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Fully automated non-native speech recognition using confusion-based acoustic model integration," in *Interspeech'2005-Eurospeech—9th European Conference on Speech Communication and Technology*, Lisboa, 2005, pp. 1369-1372.

[85] N. Minematsu, K. Osaki, and K. Hirose, "Improvement of non-native speech recognition by effectively modeling frequently observed pronunciation habits," in *EUROSPEECH 2003 - GENEVA*, Geneva, 2003, pp. 2597-2600.

[86] B. H. A. Ahmed and T. Tien-Ping, "Automatic Speech Recognition of Code Switching Speech Using 1-Best Rescoring," in *Asian Language Processing (IALP), 2012 International Conference on*, 2012, pp. 137-140.

[87] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?," *Proceedings of the IEEE*, vol. 88, pp. 1270-1278, 2000.

[88] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, vol. 42, pp. 93-108, 1// 2004.

[89] W. Kim, "Language model adaptation for automatic speech recognition and statistical machine translation," Doctoral dissertation, The Johns Hopkins University, 2004.

[90] M. Federico and N. Bertoldi, "Broadcast news LM adaptation over time," *Computer Speech & Language*, vol. 18, pp. 417-435, 10// 2004.

[91] C. Hsuan-Sheng and B. Chen, "Word Topical Mixture Models for Dynamic Language Model Adaptation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-169-IV-172.

[92] J. D. Echeverry-Correa, J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba, and J. M. Montero, "Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition," *Expert Systems with Applications*, vol. 42, pp. 101-112, 1// 2015.

[93] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," 2000.

[94] J. R. Bellegarda, "An overview of statistical language model adaptation," in *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, 2001.

[95] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," 1997.

[96] H. Nanjo and T. Kawahara, "Unsupervised language model adaptation for lecture speech recognition," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[97] Y. Liu and F. Liu, "Unsupervised language model adaptation via topic modeling based on named entity hypotheses," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4921-4924.

[98] J. M. Lucas-Cuesta, J. Ferreiros, F. Fernández-Martı´nez, J. D. Echeverry, and S. Lutfi, "On the dynamic adaptation of language models based on dialogue information," *Expert Systems with Applications*, vol. 40, pp. 1069-1085, 3// 2013.

[99] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993-1022, 2003.
[100] C. Jen-Tzung and C. Chuang-Hua, "Dirichlet Class Language Models for Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 482-495, 2011.
[101] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker, "Using information retrieval methods for language model adaptation," in *INTERSPEECH*, 2001, pp. 255-258.
[102] X. Liu, M. J. F. Gales, and P. C. Woodland, "Use of contexts in language model interpolation and adaptation," *Computer Speech & Language*, vol. 27, pp. 301-321, 1// 2013.
[103] M. Wester, "Pronunciation modeling for ASR – knowledge-based and data-derived methods," *Computer Speech & Language*, vol. 17, pp. 69-85, 1// 2003.
[104] I. Amdal and E. Fosler-Lussier, "Pronunciation variation modeling in automatic speech recognition," *Telektronikk*, vol. 99, pp. 70-82, 2003.
[105] M. Ali, M. Elshafei, M. Al-Ghamdi, and H. Al-Muhtaseb, "Arabic phonetic dictionaries for speech recognition," *Journal of Information Technology Research (JITR)*, vol. 2, pp. 67-80, 2009.
[106] K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, J. Gang, *et al.*, "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. I-344-I-347 vol.1.
[107] M. Ali, M. Elshafei, M. Al-Ghamdi, H. Al-Muhtaseb, and A. Al-Najjar, "Generation of arabic phonetic dictionaries for speech recognition," in *Innovations in Information Technology, 2008. IIT 2008. International Conference on*, 2008, pp. 59-63.
[108] H. Strik, "Pronunciation adaptation at the lexical level," in *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, 2001.
[109] T. Slobada and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, pp. 2328-2331 vol.4.
[110] E. Fosler-Lussier, S. Greenberg, and N. Morgan, "Incorporating contextual phonetics into automatic speech recognition," *Nucleus*, vol. 48993, p. 62118, 1999.
[111] H. Al-Haj, R. Hsiao, I. Lane, A. W. Black, and A. Waibel, "Pronunciation modeling for dialectal arabic speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009, pp. 525-528.
[112] F. Biadsy, N. Habash, and J. Hirschberg, "Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules," presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, Colorado, 2009.

[113] A. Ramsay, I. Alsharhan, and H. Ahmed, "Generation of a phonetic transcription for modern standard Arabic: A knowledge-based model," *Computer Speech & Language*, vol. 28, pp. 959-978, 7// 2014.
[114] F. Al-Otaibi, "speaker-dependant continuous Arabic speech recognition," M.Sc. thesis, King Saud University, 2001.
[115] S. M. Abdou, S. E. Hamid, M. Rashwan, A. Samir, O. Abdel-Hamid, M. Shahin, *et al.*, "Computer aided pronunciation learning system using speech recognition techniques," in *INTERSPEECH*, 2006, pp. 249–252.
[116] H. Soltau, G. Saon, B. Kingsbury, j. kuo, L. Mangu, D. Povey, *et al.*, "The IBM 2006 Gale Arabic ASR System," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-349-IV-352.
[117] M. Azmi, H. Tolba, S. Mahdy, and M. Fashal, "Syllable-based automatic arabic speech recognition in noisy-telephone channel," *WSEAS Transactions on Signal Processing*, vol. 4, pp. 211-220, 2008.
[118] O. Rambow, D. Chiang, M. Diab, N. Y. Habash, and S. Shareef, "Parsing arabic dialects," 2006.
[119] M. Nofal, E. Abdel Reheem, H. El Henawy, and N. Abdel Kader, "The development of acoustic models for command and control arabic speech recognition system," in *Electrical, Electronic and Computer Engineering, 2004. ICEEC '04. 2004 International Conference on*, 2004, pp. 702-705.
[120] A. Alimi and M. Ben Jemaa, "Beta fuzzy neural network application in recognition of spoken isolated Arabic words," *Control and intelligent systems*, vol. 30, pp. 47-51, 2002.
[121] S. H. El-Ramly, N. S. Abdel-Kader, and R. El-Adawi, "Neural networks used for speech recognition," in *Radio Science Conference, 2002. (NRSC 2002). Proceedings of the Nineteenth National*, 2002, pp. 200-207.
[122] H. Bahi and M. Sellami, "A hybrid approach for Arabic speech recognition," in *Computer Systems and Applications, 2003. Book of Abstracts. ACS/IEEE International Conference on computer systems and applications*, 2003, p. 107.
[123] Carnegie Mellon University. (2015, 20 Feb). *Carnegie Mellon University Statistical Language Modeling toolkit*. Available: http://www.speech.cs.cmu.edu/SLM/toolkit_documentation.html#tools
[124] CMU SphinxTrain. (2015, 1 February). Available: http://sourceforge.net/projects/cmusphinx/files/sphinxtrain/
[125] CMU Sphinx3. (2015, 1 February). Available: http://sourceforge.net/projects/cmusphinx/files/sphinx3/
[126] National Institute of Standards and Technology's. (1 January). *Speech Recognition Scoring Toolkit*. Available: http://www.itl.nist.gov/iad/mig/tools/
[127] King Saud University. (1 April). *Ayat*. Available: http://quran.ksu.edu.sa/ayat/?l=en
[128] M. W. King and P. A. Resick, "Data mining in psychological treatment research: A primer on classification and regression trees," *Journal of Consulting and Clinical Psychology*, vol. 82, pp. 895-905, 2014.
[129] M. Deza and E. Deza, "Encyclopedia of Distances," in *Encyclopedia of Distances*, ed: Springer Berlin Heidelberg, 2009, pp. 1-583.

[130] E. F. Krause, *Taxicab geometry: An adventure in non-Euclidean geometry*: Courier Corporation, 2012.
[131] E. Maor, *The Pythagorean theorem: a 4,000-year history*: Princeton University Press, 2007.
[132] N. Dunford and J. T. Schwartz, "Linear operators, vol. I," *Interscience, New York*, vol. 1963, 1958.
[133] Abduldaem Al-Kaheel. (1 April). *Secrets of Quran Miracles*. Available: http://www.kaheel7.com/book/quran-arabic-free-download.doc
[134] M. E. Ahmed, "Toward an Arabic text-to-speech system," *The Arabian Journal of Science and Engineering*, vol. 16, pp. 565-583, 1991.
[135] Mansour Alghamdi, Husni Almuhtasib, and M. Elshafei, "Arabic Phonological Rules," *King Saud University Journal: Computer Sciences and Information*, vol. 16, pp. 1-25, 2004.
[136] M. Elshafei, H. Al-Muhtaseb, and M. Al-Ghamdi, "Techniques for high quality Arabic speech synthesis," *Information Sciences*, vol. 140, pp. 255-267, 2// 2002.
[137] D. AbuZeina, W. Al-khatib, and M. Elshafei, "Small-Word Pronunciation Modeling for Arabic Speech Recognition: A Data-Driven Approach," in *Information Retrieval Technology*. vol. 7097, M. Salem, K. Shaalan, F. Oroumchian, A. Shakery, and H. Khelalfa, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 529-537.
[138] D. AbuZeina, W. Al-Khatib, M. Elshafei, and H. Al-Muhtaseb, "Toward enhanced Arabic speech recognition using part of speech tagging," *International Journal of Speech Technology*, vol. 14, pp. 419-426, 2011/12/01 2011.