AUTOMATIC IDENTIFICATION OF ARABIC DIALECTS USING HIDDEN MARKOV MODELS

by

Fawzi S Alorifi

B.S. E.E., King Saud University, Riyadh, Saudi Arabia, 1989

M.S. E.E., University of Pittsburgh, Pittsburgh, USA, 1998

Submitted to the Graduate Faculty of
the Swanson School of Engineering in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Fawzi S Alorifi

It was defended on

June 18, 2008

and approved by

Dr. Amro A. El-Jaroudi, Associate Professor, Electrical and Computer Engineering Department

Dr. J. Robert Boston, Professor, Electrical and Computer Engineering Department

Dr. Luis F. Chaparro, Associate Professor, Electrical and Computer Engineering Department

Dr. Ching-Chung Li, Professor, Electrical and Computer Engineering Department

Dr. John D. Durrant, Professor, Communication Science and Disorders Department, School of Health and Rehabilitation Sciences

Dr. Susan Shaiman, Associate Professor, Communication Science and Disorders Department, School of Health and Rehabilitation Sciences

Dissertation Director: Dr. Amro A. El-Jaroudi, Associate Professor, Electrical and Computer Engineering Department

ABSTRACT

AUTOMATIC IDENTIFICATION OF ARABIC DIALECTS USING HIDDEN MARKOV MODELS

Fawzi S Alorifi, PhD

University of Pittsburgh, 2008

The Arabic language has many different dialects, they must be identified before Automatic Speech Recognition can take place. This thesis examines the difficult task of properly identifying various Arabic dialects. We present a novel design of an Arabic dialect identification system using Hidden Markov Models (HMM). Due to the similarities and the differences among Arabic dialects, we build a ergodic HMM that has two types of states; one of them represents the common sounds across Arabic dialects, while the other represents the unique sounds of the specific dialect. We tie the common states across all models since they share the same sounds. We focus only on two major dialects: Egyptian and Gulf. An improved initialization process is used to achieve better Arabic dialect identification. Moreover, we utilize many different combinations of speech features related to MFCC such as time derivatives, energy, and the Shifted Delta Cepstra in training and testing the system. We present a detailed comparison of the performance of our Arabic dialect identification system using the different combinations. The best result of the Arabic dialect identification system is 96.67% correct identification.

keywords: Language Identification, Gaussian Mixture Models, GMM, Egyptian Dialect, Gulf Dialect, Arabic Dialect Database, HMM Initializations, Jackknifing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

My parents deserve my sincere appreciation and good prayers for their great and precious support, caring, and guidance. Without their encouragement and support, this work and all of my successes would not have been a reality.

I would like to thank my advisor Dr. Amro A. El-Jaroudi for the guidance and support he has provided me throughout my Ph.D. studies. I would like also to express my appreciation to all of the members serving on my dissertation committee for their time and invaluable feedback.

I would like to express my gratitude to the Saudi Arabian Government for sponsoring my graduate studies.

Also I am grateful to my brothers, friends and colleagues Dr. Fahad Alorifi, Dr. Ahmed Alorifi, Nazeeh Alothmany, Basil AsSadhan, Dr. Saeed Aldosari, Mansour Alawaji, Hesham Albesher, Talal Alkhateeb, Massimo Cenciarini, and Dr. Arash Mahboobin.

Last but not the least, I would like to thank my friend and wife, Layla, for her endless sacrifices, support, and sharing with me the ups and downs. Her patience in managing our home and raising our children is an unforgettable effort. Without her sacrifices, I am sure that this work would not exist. Also, I dedicate this dissertation to my lovely kids Haifa, Nasser, Talal, Salma, Yasmeen, and Randa.

1.0 INTRODUCTION

1.1 MOTIVATION AND SCOPE

In general, Automatic Speech Recognition for English and other languages has been the subject of much research in the last forty years. Some of its problems have been solved successfully, while others are still under investigation. In contrast, speech recognition for the Arabic language only has been researched since the 1980's [3].

By and large, Arabic language research has been growing very slowly in comparison to English language research. This slow growth is mainly due to a lack of modern studies on the acoustic-phonetic nature of the Arabic language and to the inherent difficulties in speech recognition. Additionally, there is no standardized database of Modern Standard Arabic (MSA) language in general, nor are there many modern studies that examine Arabic dialects for speech identification purposes [3]. Past studies examining Arabic dialects for speech identification were written purely from linguistic points of view, explored antiquated dialects, or focused on specific dialects; very few examined any modern forms of Arabic.

Most Automatic Speech Recognition (ASR) systems are based on speech selected from the standard form of a language; however, dialects of a language can be quite different from the standard form. Since dialects can be structurally very different, it is essential to develop methods to automatically identify them [4].

The performance of an ASR system is affected by how the training data is matched to the test data. Having a mismatch in dialect between training and testing will result in poor recognition performance [5]. Consequently, it is important that the dialect of a speaker be identified so that models based on the appropriate training data be used during recognition. Moreover, many dialects (especially Arabic) have a different linguistic structure, so it is essential to identify the dialect to use the appropriate vocabulary and grammar during speech recognition.

These problems motivate us to work on Automatic identification for Arabic dialects. Most of the ASR systems for Arabic are based on the MSA language, but in reality, most people speak a regional dialects. Therefore, identifying the Arabic dialect from the input speech will help the ASR of Arabic to have optimal performance.

1.2 THESIS OBJECTIVES

In this research, we present a system for automatic Arabic dialect identification. This system is designed based on the characteristics, the similarities and the differences, of Arabic dialects. The Arabic dialect identification system is based upon the Ergodic Hidden Markov Model (EHMM). Our system takes advantage of the similarities and differences between Arabic dialects and utilizes models with two separate types of states: one corresponding to the unique sounds in each Arabic dialect, and the other corresponding to the common sounds in all Arabic dialects. To ensure that common states remain the same in all models, the common states are tied together.

This research concentrated on the following topics:

1. The structure of Hidden Markov Models of the dialect models.
2. The Arabic dialect database.
3. The optimal speech feature vectors based on the performance of the Arabic dialect identification system.
4. The initialization of the Hidden Markov Model.

1.3 DISSERTATION ORGANIZATION

In chapter 2 we provide background information about Arabic dialects including related research. Also, in the same chapter, we categorize the Arabic dialects into five major dialects: the Iraqi Dialect, the Gulf Dialect, the Levant Dialect, the Egyptian Dialect, the Maghreb Dialect. Chapter 3 provides relevant theoretical information regarding speech processing, Gaussian mixture models (GMM), and Hidden Markov models. In chapter 4, we discuss the speech database used in the study. Three training databases are used throughout the work, an unbalanced training database and two balanced training databases. The Arabic dialect identification system is also presented in this chapter. In the final section of chapter 4, results for the identification system are presented. Many configured Hidden Markov models, trained and tested, are provided with illustrations throughout this chapter. Chapter 5 presents modifications and improvements to the Arabic dialect identification system. Moreover, different speech features are used in training to find the best features for the system to perform well. Finally, chapter 6 summarizes the contributions of this work and explores future research possibilities.

2.0 ARABIC DIALECTS BACKGROUND

In this chapter we investigate and review the Arabic dialects featured in this study and review pervious research on Dialect Identification.

2.1 ARABIC DIALECTS

The Arabic Language is one of the oldest living languages in the world. The bulk of classical Islamic literature was written in Classical Arabic (CA), and the Holy Qur'an was revealed to Prophet Mohammed, peace be upon him, in the Classical Arabic language. Standard Arabic is the mother (spoken) tongue for more than 200 million people living in the vast geographical area known as the Arab world, which includes countries such as Syria, Jordan, Egypt, Saudi Arabia, Morocco, and Sudan [6, 3]. More information regarding the Arabic language in general can be found In Appendix A. Figure 2.1 illustrates the nations where Arabic is spoken as a mother tongue.

Modern Standard Arabic (MSA) dialect lacks the extreme formality of Classical Arabic, and it is considered the standard dialect of all Arabic speaking nations. Like many languages, Arabic has branched off into numerous colloquial variations. MSA's pronunciation and lexicon can be drastically altered when mixed with local dialects; the use of both colloquial and Modern Standard Arabic in one's daily activities can be described as Diglossia, a condition in which two varieties of an identical language are used in different conditions within a community, often by the same speakers [7, 8, 9]. MSA is used in written texts, formal speeches, sermons, and news broadcasts, while colloquial Arabic is used between family or friends and

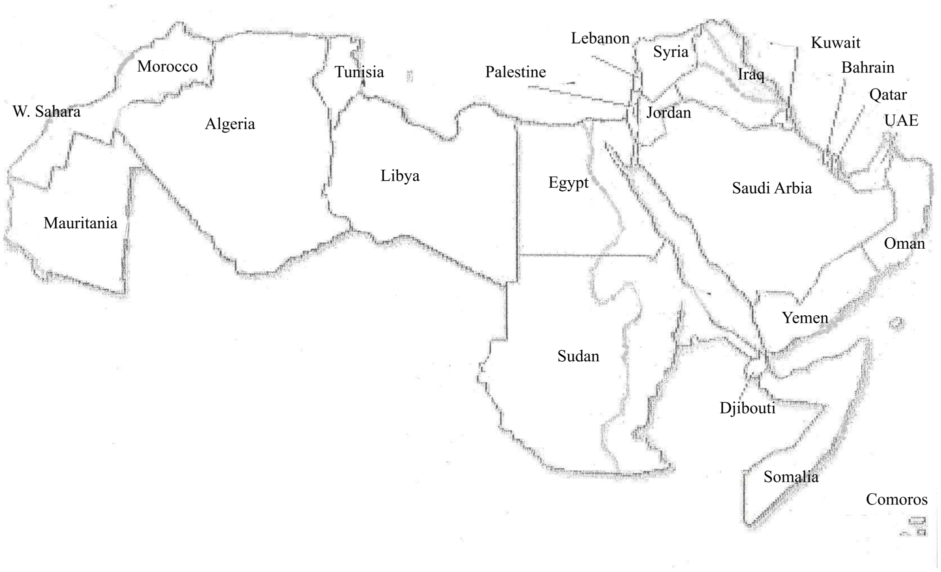

Figure 2.1: Arab Countries Map.

on TV and radio soap operas. Colloquial Arabic does not only differ among nations, but also among regions within the same country [7, 8]. For example, one can find 200 different dialects in the Arabian Peninsula alone [10, 11]. Dialects can be similar and different at the same time; while the Moroccan and Egyptian dialects share the same grammar, phonology, and lexicon, they remain to be very distinct dialects [8]. Differences in individual dialects can be categorized into three general types:

- **Eastern vs. Western Dialects:** The foundations of the Arabic language were laid in the Arabian Peninsula. Arabic continued to undergo major transformations until the third decade of the seventh century, when Arabic began to spread alongside Islam to the north, east, and west [7]. As the language spread under these conditions, it caused eastern and western dialects to become more alike, though they still remained unique in their own ways. The further a nation was from the Arabian Peninsula, the more distinct and different its dialect; because of the great distance between the eastern and western Arabic nations, a linguistic dichotomy continues to exist between them [8]. Countries that speak an Eastern form of Arabic include the Arabian Peninsula (Saudi Arabia, Kuwait, Qatar, Bahrain, United Arab Emirates, Oman, Yemen) along with Iraq, Syria, Lebanon, Palestine, Jordan, Egypt, and Sudan. Speakers of western dialects can be found in Libya, Tunisia, Algeria, Morocco, and Mauritania. While there are many differences between the Eastern and Western dialects, the primary difference is phonetic in nature; speakers of Western dialects tend to drop many short vowels, and reduce the length of many long vowels. In contrast, speakers of Eastern dialects tend to maintain classical vowels. Additionally, speakers of Western dialects make no phonemic distinction between *Seen* and *Saad* and between *Zaa* and *Dhaa*, while Eastern speakers do. Eastern and Western dialects differ morphologically, syntactically, and lexically as well [8]. Regarding morphological differences, speakers of western dialects tend use *n* in both singular and plural cases, while eastern dialects utilize *n* only in the plural case. For example, "*I/We write*" is stated in western dialects as "*naktab-naktabu*" respectively, while in eastern dialects it is stated "*'aktib-niktib*" [7, 8].

- **Bedouin vs. Urban Dialects** Historically, Arabic has been a collection of various dialects, some originating from bedouin life, and others from urban, sedentary life [8]. Urban dialects are mixtures of Classical Arabic and a foreign substratum, while bedouin dialects are more isolated from foreign influence, helping them to resist drastic linguistic evolution, and retain features of Classical Arabic. For example, a prominent feature of bedouin dialects is voicing the normally unvoiced uvular stop *Qaaf* (q), causing it to become either a voiced and fronted *Qaaf* (G) or a voiced, fronted, and affricated *Jiim* (dz) [8, 11]. Another characteristic of bedouin dialects is that gender is generalized in plural pronouns, more so than in urban dialects. Additionally, many bedouin lexemes such as *husaam* (sword) originate from Classical Arabic, though they are not used in urban dialects. Identical lexemes may have entirely different meanings between bedouin and urban dialects, such as *dahraj*, which means "see" in bedouin Arabic but to "roll something" in Palestinian dialects [8]. A final distinction between bedouin and urban dialects involves syllable structure.

- **Religious Dialects: Muslim, Christian, and Jewish Dialects** In addition to regional dialects, Arabic also contains some linguistic variation due to religion, most notably Islam, Christianity, and Judaism. Also, sects of different religions exhibit varying dialects; for example, Lebanese Shiites, Sunnis, Druze, and Maronites all speak unique dialects [8].

Since there are numerous types of Arabic dialects, time constraints necessitate that the research focus only on dialects that belong to certain social, ethnic, and religious groups, since these factors tend to be associated with the most numerous new or systematic changes.

Thus, five regional dialects were identified as core dialects; Figure 2.2 provides an illustration of where each dialect is spoken by region:

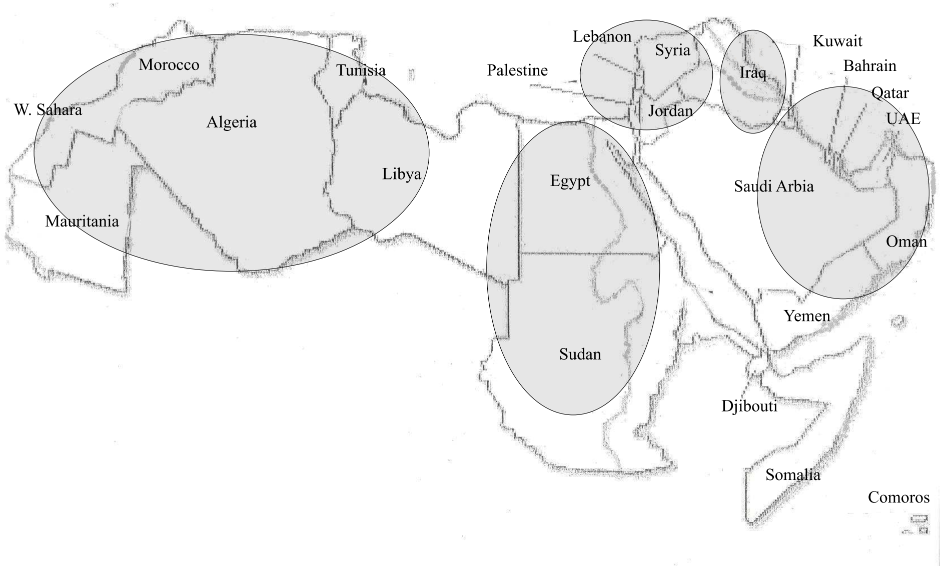

Figure 2.2: Arabic Regional Dialects.

- **The Iraqi Dialect** The Iraqi Dialect is spoken only in Iraq. A characteristic of the Iraqi dialect is that the phoneme *Qaaf* can be changed to a voiced *Qaaf* (G) or unvoiced *Qaaf* (q).

- **The Gulf Dialect** The Gulf dialect is spoken around the shores of the Arabian Gulf (Persian Gulf) which includes Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, and Oman. In these dialects, the phoneme *Jiim* (dz) can be changed to *Yaa* (j). Additionally, *Qaaf* (q) can be substituted with a voiced *Qaaf* (G).

- **The Levant Dialect** The Levantine dialect is spoken by Arabs near the Mediterranean east coast, including countries such as Jordan, Syria, Lebanon, and Palestine. This dialect can be divided into six sub dialects:
  Lebanese dialect
  Central Syrian
  Northern Syrian
  Rural Palestinian
  Urban Palestinian
  Bedouin Palestinian
  Commonly changed phonemes in the Levant dialects occur only in urban forms, where the interdentals (i.e. *Dhaa*, *Thaa*, and *Thaal*) are changed to *Taa*, *Daal*, and *Dhaad*. Lastly, *Qaaf* (q) is changed to *Qaaf* (q) with a glottal stop.

- **The Egyptian Dialect** As its name suggests, Egyptian Arabic is spoken in Egypt. However, since Egypt is one of the most heavily populated Arab nations, the dialect has become extremely widespread and well-known. This is due largely to Egypt's large contribution to Arab filmmaking and television production, which has inadvertently made the Egyptian dialect understood throughout the Arab world. Consequently, the Egyptian dialect is frequently taught in many American and European schools due to its popularity and because it, like MSA, can be understood in most of the Arab world. Phonemes modified in the Egyptian dialect include the interdentals (i.e. *Dhaa*, *Thaa*, and *Thaal*), which are changed to *Taa*, *Daal*, and *Dhaad*. The Egyptian dialect also substitutes a voiced *Qaaf* (G) for the letter *Jiim* (dz), example *gabl* for *jabl* (mountain) or *gamiil* for *jamiil* (beautiful). Moreover, *Qaaf* (q) is changed to *Qaaf* (q) with a glottal stop.

Table 2.1: Spoken Languages in selected countries.

Country	1^{st} **Language**	2^{nd} **Language**	3^{rd} **Language**
Saudi Arabia	Colloquial Gulf Arabic	Standard Arabic	English
United Arab Emirates	Colloquial Gulf Arabic	Standard Arabic	English
Egypt	Colloquial Egyptian Arabic	Standard Arabic	English
Morocco	Colloquial Maghreb Arabic	Standard Arabic	French
Tunisia	Colloquial Maghreb Arabic	Standard Arabic	French
Palestine, Israel	Colloquial Levantine Arabic	Standard Arabic	Hebrew

- **The Maghreb Dialect** Magherb refers to an Arab geographical region including Morocco, Tunisia, Algeria, and Western Libya. The Maghreb dialect is a spoken language in the aforementioned regions, and labeled by the majority of its speakers as *Derija*, meaning "dialect". Since the Maghreb region was colonized by France and Spain, its dialect combines many French and Spanish root words with Arabic suffixes to form words. Since this form of Arabic is rarely written, it is less static, and changes frequently. The Maghreb dialect's phonemes differ in that speakers make no distinction between short and long vowels.

As was the case in Maghreb, other Arabic dialects were affected by foreign languages, usually due to colonization. For example, Table 2.1 shows the languages spoken by native speakers in selected countries [12, 13].

These dialects share similarities and differences from a phonetic point of view. For example, major phonemes that change among these dialects are vowels (long and short), interdentals (i.e. *Dhaa*, *Thaa*, and *Thaal*), *Qaaf* , *Kaaf* , and *Jiim*. Listed below are dialects that are similar or different with regard to the phonetic characteristics discussed above [8, 14].

- **Phoneme *Jiim* (dz)**
  - Iraqi, Damascus-Levantine, and Maghreb dialects use a similar (dz).
  - Gulf dialects use *Yaa* (j).
  - The Egyptian dialect changes it to a voiced *Qaaf* (G).
- **Phoneme *Qaaf* (q)**
  - Iraqi and Gulf dialects both use a voiced *Qaaf* (G).
  - Maghreb speakers use *Qaaf* (q).
  - Levantine and Egyptian dialects use an identical *Qaaf* (q) with glottal stop.
- **Phoeme *Kaaf* (k)**
  - Maghreb, Levantine, and Egyptian dialects retain (k).
  - Gulf and Iraqi dialects make (k) an affricate.
- **Interdentals (*Dhaa*, *Thaa*, and *Thaal*)**
  - Iraqi and Gulf dialects retain interdentals.
  - Egyptian, Levantine dialects replace interdentals with *Taa*, *Daal*, and *Dhaad*
  - Maghreb dialect replace interdentals with homorganic stops.
- **Long Vowels [a: e: i: o: u:]**
  - Iraqi, Levantine, Gulf, and Egyptian dialects retain long vowels.
  - The Maghreb dialect makes no distinction between long or short vowels; it has only six vowels (3 stable and 3 variable).
- **Short Vowels [i a u]**
  - Iraqi dialect uses a /o/.
  - Gulf dialects retain all three short vowels.
  - Egyptian Arabic shifts /a/ for /i/.
  - Levantine Arabic retains short vowels.
  - The vowel structure of the Maghreb dialect is completely different as described above.

2.2 RELATED RESEARCH

What follows is a literature review of previous work examining dialect identification. In reviewing these works we focus on three factors:

- Size of the research database.
- Type of recognition system utilized.
- Results obtained.

Barakat's [4] research utilized perceptual Arabic dialect identification; speech data was comprised of two major dialects: Eastern (Middle East) and Western (Maghreb). Data was collected by requesting respondents to spontaneously describe various pictures; 12 participants originated from 6 different Arabic speaking countries: Morocco, Algeria, Tunisia, Syria, Lebanon, and Jordan. The respondents' data was recorded on tape, and then used as stimulus material for 18 different subjects, who were all native speakers from the same six nations; the second group of respondents were asked to listen to these recordings and determine the origin of the speakers. Of the 18 subjects in the second group, 97% correctly identified the Maghrab dialect; likewise, 99% correctly identified Eastern dialects.

In another study by Barakat et al. [15], prosody, or the rhythm of one's speech, was used as a method of discriminating between different Arabic dialects. Speech data in this study consisted of participants from four countries: Morocco, Algeria, Syria, and Jordan. Both natural and synthetic speech was used as stimuli for 38 adult listeners, who were divided into two groups: native speakers from Maghrab countries, and those with limited knowledge of Arabic. For the natural speech samples, 97% of native, and 56% of non-native speakers correctly identified the speaker's dialect. However, when participants listened to the synthetic speech samples only 58% of native speakers, and 49% of non-native speakers correctly identified the sample's dialect.

Marc Zissman et. al. [16] designed a system to distinguish between Cuban and Peruvian dialects of Spanish. In order to carry out this task, the researchers built a corpus consisting of 219 speakers who each spoke for 20-30 minutes each. Their system utilized techniques of phone recognition, along with language modeling (PRLM). In this system, training messages in each dialect were tokenized by a s single-language phone recognizer; the resulting symbol sequence associated with each of the training messages is analyzed, and an n-gram probability distribution language model was estimated for each dialect. The system used an English phone recognizer trained on the TIMIT corpus because it was a labeled corpus. Using speech from 143 Cuban and Peruvian speakers, the authors divided speakers into three groups according to how typical a dialect was to his or her native region. Two groups were typical dialects, consisting of speech from 119 different speakers, while the third group consisted of atypical dialects, comprised of speech from 24 different speakers. Experiments were conducted on all three groups. During training, three minutes of spontaneous Spanish speech from each speaker in the Cuban training set were processed by the English phone recognizer, and the Cuban language model statistics were computed. The same thing was repeated for the Peruvian speakers. With the two language models in hand, test-speakers' spontaneous speech was processed and the Dialect Identification, DID, decision was computed. The test utterances were three minutes long. The dialect error rate was 16%.

Itahashi Shuichi et. al. [17] attempted to classify dialects according to the speech fundamental frequency (F0) of the speaker. The researchers used F0 to determine which features were most useful in spoken language and dialect classification. After F0 was determined, speech power was used to detect voiced intervals; the fundamental frequency contour was approximated and represented by polygonal lines for each voiced interval. Finally, statistical parameters were calculated from the polygonal lines. To perform language discrimination, samples were collected from four male speakers for each of the following six languages: Japanese, Chinese, Korean, English, French, and German. In the dialect analysis, 12 Japanese dialects were examined from 2 or 4 different male respondents each. Both closed and open experiments were conducted; in closed experiments, the same sample was used for training and testing, while in the open one, samples used in training were not in the final examination. In the closed and open experiments, the dialect identification success rate was 89.1% and 75% respectively. For the dialect identification of the 12 Japanese dialects, the result was 79.2% without indicating if it was obtained from open or closed experiment.

Itahashi Shuichi et. al. [18] used their previous method in a second study to classify languages and dialects based on speech fundamental frequency. In this study, the same six languages were used, however the number of Japanese dialects was increased to fourteen dialects. The samples consisted of five males for each languages and one male for each dialect. The results for language were 100% for the closed experiment and 80% for the open experiments. For the dialect, the results were 94% for the closed experiment and 61.9% for the open experiment.

From the literature review above, different methods were used to identify dialects or languages ranging from asking subjects to listen to dialects to automatic dialect identification systems. Also, the methods used prosody, fundamental frequency, or phone recognition along with langauge modeling.

Our work addresses the problem of automatic Arabic dialect identification. We build a database of audio examples for two Arabic dialects. We develop a novel modeling approach that takes into account the differences and similarities between the dialects. We use Hidden Markov Modeling methods to build the dialects models and for dialect identification. We also optimize our approach by investigating the effects of model structure, model dimension, model initialization, and the model acoustical feature set, among others. Our efforts to improve the performance of the dialect identification system achieve a maximum correct identification rate of 96.7% on our test database.

3.0 SPEECH PROCESSING AND IDENTIFICATION SYSTEM

In this chapter we present the theory of the system used in our research. Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) are often used in some identification systems of language, dialect, accent, or speaker [19, 16, 20]. Since this research utilizes both techniques, this chapter will explore these methods briefly, in addition to feature extraction.

The block diagram shown in Figure 3.1 illustrates the main parts of a speech recognition system [1]. Each block in the Figure is described briefly below and in more detail later in this chapter.

Signal Processing In this process, the speech signal is converted to a set of feature vectors.

Acoustic Models The representation of knowledge about acoustic, phonetics, and the speaker variability are included in the models. Hidden Markov Models are the foundation for acoustic phonetics models. The acoustic models are modified during training to ensure that system performance is optimized.

Language Models The knowledge of the system about what words are likely to appear together, in what sequence, and what the possible words are.

Recognition Algorithm (Decoder) In this process, the decoder matches the input feature vectors to the acoustic models and language models to find the most likely word sequence.

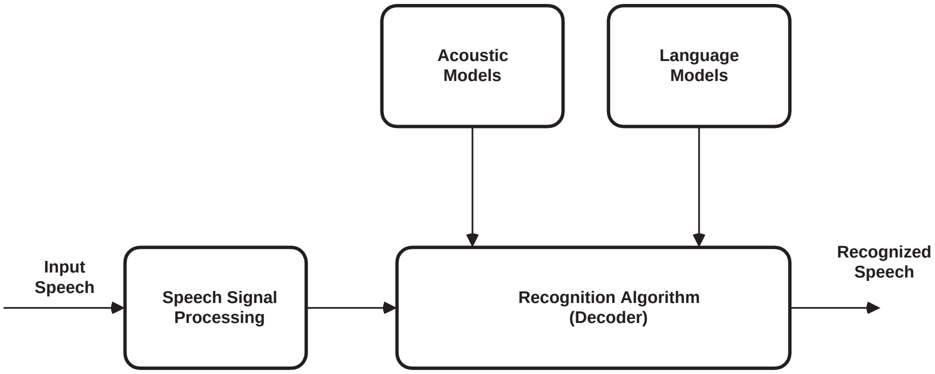

Figure 3.1: Structure of a Speech Recognition System [1].

3.1 FEATURE EXTRACTION

Feature extraction is a fundamental part of any speech recognition or identification system. Feature extraction is a stage where speech is first transformed into a set of speech frames. Since the speech signal is non stationary, the speech frames should be small enough to be assumed stationary; moreover, the speech frames should be long enough to contain relevant information about the speech. Speech must be pre-processed before features can be extracted. Pre-processing includes a stage of preemphasis, where a digitized speech signal is filtered by a first-order Finite Impulse Response (FIR) filter in order to flatten the spectrum of the signal and increase the relative energy of high frequencies [21].

Typically, the speech frame is between 20 to 30 milliseconds long, while the overlap and frame rate are 10 milliseconds, 100 frames per second, respectively [21]. A window is then applied to each resulting speech frame to minimize discontinuities at the beginning and end of each frame. The most common window used is the Hamming window whose length is equal to that of the speech frame [21, 22]. The front end used in our work is a generic front end in many speech recognition systems and the frame length is appropriate for the cepstral features to be extracted. However, some features, such as pitch, may be important in dialects of other languages. In such cases, other front end parameters may be appropriate.

An utterance is represented as a sequence of feature vectors. There are many representations of the features, however the cepstrum is one of the most widely used in speech recognition [23]. The cepstrum is defined as the inverse Fourier transform of the logarithmic of the short time spectrum. Lower order cepstrum coefficients represent the vocal tract impulse response. The cepstrum of a signal $x(n)$ is computed using the following steps [24, 25]

1. Pre-emphasising the signal
2. Windowing the signal;
3. Computing the Fast Fourier Transform (FFT);
4. Taking the magnitude the spectrum;
5. Calculating the log;
6. Calculating the inverse FFT.

Equation 3.1 explains the previous steps where $c(n)$ is the cepstrum, while $X(e^{jw})$ is the Fourier transform of the pre emphasized and windowed signal $x(n)$

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} log|X(e^{jw})| \, e^{jwn} \, dw \tag{3.1}$$

Mel-Frequency Cepstral Coefficients (MFCC) are the most popular feature in speech identification systems in deference to the place-pitch mechanism originating along the hearing origin [1, 26]. To compute MFCC, filter banks of triangular shape are formed. They are equally spaced along the Mel-scale defined by

$$Mel(f) = 2595 \; log_{10} \; (1 + \frac{f}{700}) \tag{3.2}$$

A Fourier transform is applied to the framed speech data and the magnitude squared spectrum is computed [27, 28]. The spectrum is multiplied by the corresponding triangular filter gain and the results are accumulated for each filter. These results hold a weighted sum representing the spectral magnitude squared in that filter bank. Then, using Discrete Cosine Transform (DCT), the cepstral coefficients are computed from the log filter bank magnitudes (m_j) [25, 28].

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N} (j - 0.5)\right) \tag{3.3}$$

The MFCC calculation takes into account the auditory characteristics of the human ear, which resolves frequencies non-linearly across the audio spectrum; it has been suggested that the development of a similar non-linear method would improve recognition performance. The time derivatives of the MFCC are usually appended to the feature vector in order to capture the dynamics of speech. To make the cepstral features more robust, Cepstral Mean Normalization (CMN) is used [28, 1]. Since time-domain convolutional distortions such as reverberations and differences in sound quality due to different types of microphones become additive offsets in the cepstral domain, subtracting the noise component from the distorted speech helps to provide clean speech features. CMN approximates all time-invariant frequency distortions and the convolutional noise component by means of the cepstral mean vector, where $\overline{\mathbf{c}}$ is given by [1, 29]

$$\overline{\mathbf{c}} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{c}_t \tag{3.4}$$

CMN subtracts $\overline{\mathbf{c}}$ from $\mathbf{c}_t$ to obtain the normalized cepstrum vector $\widehat{\mathbf{c}}$:

$$\widehat{\mathbf{c}} = \mathbf{c}_t - \overline{\mathbf{c}} \tag{3.5}$$

Removing this mean from the ceptsral coefficients will make the identification system more reliable [1, 29]. Next, The recognition algorithm, as in Figure 3.1 is discussed.

3.2 RECOGNITION ALGORITHM (DECODER)

As shown in Figure 3.1, after the speech has been processed and converted into sequences of acoustic vectors $\mathbf{Y} = \mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_T}$, the decoder determines the most probable word sequence $\widehat{W}$ given by the observed acoustic signal $\mathbf{Y}$. Using Bayes's rule, $P(W|\mathbf{Y})$ can be divided into two components:

$$\widehat{W} = \underset{W}{\text{argmax}} \quad P(W|\mathbf{Y}) = \underset{W}{\text{argmax}} \quad \frac{P(W)P(\mathbf{Y}|W)}{P(\mathbf{Y})} \tag{3.6}$$

This means that to find the most likely word sequence $\widehat{W}$, the word sequence which maximizes the product $P(W)P(\mathbf{Y}|W)$ must be found. The second term, $P(\mathbf{Y}|W)$ represents the distribution of acoustic features for a given word. Calculation of $P(\mathbf{Y}|W)$ requires the design of suitable sub-word models. Hidden Markov Models (HMMs) are used in this step [30].

The first term $P(W)$ represents the language. The purpose of the language model is to provide a process of estimating the probability of a given word W_k in an utterance considering the preceding words $W = W_1, W_2, \cdots, W_{k-1}$. A simple way to find these probabilities is to use an "N-gram" which assumes that W_k depends only upon the preceding $N-1$ words, otherwise, the calculation of such probability will be prohibitive [30] as in Equation 3.7. The most commonly used N-grams are bigrams and trigrams where $N = 2$ and 3, respectively [30].

$$P(w_k|W_l^{k-1}) = P(w_k|W_{k-N+1}^{k-1}) \tag{3.7}$$

3.3 GAUSSIAN MIXTURE MODELS (GMM)

A Gaussian mixture density $P(\mathbf{x}|\lambda)$ is defined as the weighted sum of m unimodal Gaussian densities and can be represented by the following equation [27, 31]

$$P(\mathbf{x}|\lambda) = \sum_{i=1}^{m} w_i \; p_i(\mathbf{x}) \tag{3.8}$$

The density $p_i(\mathbf{x})$ is parameterized by a mean vector $\boldsymbol{\mu_i}$ of dimension $D{\times}\mathit{1}$ and covariance matrix Σ_i of dimension $D{\times}D$ where D is the dimension of vector $\mathbf{x}$.

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\,(\Sigma_i)^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \tag{3.9}$$

The mixture weights w_i satisfy the constraint $\sum_{i=1}^{m} w_i = 1$.
The parameters, mean vectors, covariance matrices, and mixture weights are represented by the following equation [27, 31]

$$\lambda = [w_i, \boldsymbol{\mu}_i, \Sigma_i] \tag{3.10}$$

where $i = 1, \cdots, m$

There are a number of techniques for estimating the parameters of the GMM, the most popular one being Maximum Likelihood (ML) estimation. The goal of ML is to find the model λ parameters that maximize the likelihood of the GMM. For training data, $\mathbf{X} = (\mathbf{x}_i, \cdots, \mathbf{x}_T)$ the GMM likelihood is as follows

$$P(X|\lambda) = \prod_{t=1}^{T} P(\mathbf{x_t}|\lambda) \tag{3.11}$$

However, this equation is a nonlinear function of the parameters λ and direct maximization is not feasible [27, 31]. Yet, a local maximum can be reached by using the Expectation-Maximization (EM) algorithm [32]. The EM algorithm starts with initial model λ, then obtains a new model $\overline{\lambda}$ so that

$$P(X|\overline{\lambda}) \geq P(X|\lambda) \tag{3.12}$$

Then in the next iteration, the new model $\overline{\lambda}$ is used to produce a new one, repeating until convergence is reached. To make sure that the likelihood of the estimated model is increased monotonically, re-estimation equations are used in each EM iteration [27, 31]. These equations are:

Mixture Weights

$$\overline{w_i} = \frac{1}{T} \sum_{t=1}^{T} p(i|\mathbf{x}_t, \lambda) \tag{3.13}$$

Means

$$\overline{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\mathbf{x}_t, \lambda)\mathbf{x}_t}{\sum_{t=1}^{T} p(i|\mathbf{x}_t, \lambda)} \tag{3.14}$$

Covariance

$$\overline{\Sigma}_i = \frac{\displaystyle\sum_{t=1}^{T} p(i|\mathbf{x}_t, \lambda)\,(\mathbf{x}_t - \overline{\boldsymbol{\mu}}_i)\,(\mathbf{x}_t - \overline{\boldsymbol{\mu}}_i)'}{\displaystyle\sum_{t=1}^{T} p(i|\mathbf{x}_t, \lambda)} \tag{3.15}$$

where the posteriori probability of the i^{th} mixture is given by

$$p(i|\mathbf{x}_t, \lambda) = \frac{w_i p_i(\mathbf{x}_t)}{\displaystyle\sum_{k=1}^{m} w_k p_k(\mathbf{x}_t)} \tag{3.16}$$

When the covariance matrix Σ_i is assumed to be diagonal, the re-estimation equation for Σ_i can be simplified to [27, 31]

$$\overline{\rho}_{i,j}^2 = \frac{\displaystyle\sum_{t=1}^{T} p(i|\mathbf{x}_t, \lambda)\, x_j^2}{\displaystyle\sum_{t=1}^{T} p(i|\mathbf{x}_t,\ \lambda)} - \overline{\mu}_{i,j}^2 \tag{3.17}$$

for $1 \leq j \leq D$

where x_j and $\overline{\mu}_{i,j}$ are the j^{th} element of the vector $\mathbf{x}_t$, and $\boldsymbol{\mu}_i$; $\rho_{i,j}$ is the j^{th} diagonal element of the diagonal matrix $\overline{\Sigma}_i$.

3.4 HIDDEN MARKOV MODELS

Hidden Markov Models (HMMs) have become the most popular tool for building speech recognition systems. In this section, we present the basic concepts of HMMs. The HMM is defined as a “doubly stochastic process with an underlying stochastic process that is not observed (i.e. hidden), but can be observed through another set of stochastic processes that produce the sequence of the observed symbols” [33].

A Hidden Markov Model is basically a markov chain where the output observation is a random variable X generated according to an output probabilistic function associated with each state. There is no correspondence between the observation sequence and state sequence, meaning that one cannot determine state sequence from the observation sequence; therefore, the state sequence is not observable, or “hidden,” as its name suggests. An HMM can be characterized by the following elements [22]:

1. The number of states in the model: N. The individual states are denoted by $\mathbf{S} = \{S_1, S_2, \cdots, S_N\}$, and the state at time t is q_t.
2. The number of observation symbols per state, M. The set of symbol observations are denoted by $\mathbf{V} = \{v_1, v_2, \ldots, v_M\}$. Observations can be also continuous.
3. A set of state transition probabilities $A = a_{ij}$

   $$a_{ij} = P[q_{t+1} = S_j |\ q_t = S_i] \tag{3.18}$$

   for $1 \leq i, j \leq N$.
4. A probability distribution in each of the states $B = b_j(k)$ in which

   $$b_j(k) = P[v_k \,\text{at t}|\ q_t = S_j] \tag{3.19}$$

   where $\ 1 \leq k \leq M \quad 1 \leq j \leq N$.
5. The initial state distribution $\pi = \pi_i$ in which

   $$\pi_i = P[q_1 = S_i] \tag{3.20}$$

   where $1 \leq i \leq N$.

Therefore, we can use the compact notation to represent the HMM

$$\lambda = (A, B, \pi) \tag{3.21}$$

We have discussed a discrete HMM where the set of observations is discrete. However, when the set of observations is continuous, then we will have to use a continuous probability density function instead of a set of discrete probabilities. Typically, the probability density is approximated by a weighted sum of M Gaussian distributions [34].

$$b_j(o) = \sum_{k=1}^{M} w_{jk} \ N(o, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \tag{3.22}$$

where $\quad 1 \leq j \leq N \quad$ and

$N()$ = Gaussian distribution function

w_{jk} = Weighting coefficients

$\boldsymbol{\mu}_{jk}$ = Mean vector

$\boldsymbol{\Sigma}_{jk}$ = Covariance matrix

w_{jk} should satisfy the stochastic constraints

$$\sum_{k=1}^{M} w_{jk} = 1$$

where $w_{jk} \geq 0, \quad 1 \leq j \leq N, \quad 1 \leq k \leq M.$

Hence, the complete parameter notation set of an HMM with continuous probability distribution is

$$\lambda = (A, w_{jk}, \ \boldsymbol{\mu}_{jk}, \ \boldsymbol{\Sigma}_{jk}, \pi) \tag{3.23}$$

There are three basic algorithms associated with Hidden Markov Models:

1. The *forward algorithm*
2. The *Viterbi algorithm*
3. The *forward-backward algorithm*

The forward algorithm is used to compute the evaluation problem, i.e., computing $P(O/\lambda)$, the probability of model λ emitting observation sequence $O = o_1, \cdots, o_T$. Computing $P(O/\lambda)$ directly can be computationally infeasible (on the order of $2TN^T$ calculations), so using the forward algorithm is more efficient with calculations on the order of N^2T.

The Viterbi algorithm is used to solve the decoding problem i.e. finding the state sequence that maximizes the likelihood of the observations. The Viterbi algorithm is used to find the best path though the states $q = (q_1, \cdots, q_T)$ of an HMM model for a given observation sequence.

Lastly, the forward-backward algorithm is used for training purposes to optimize the model parameters $\lambda = (A, B, \pi)$. This algorithm is also known as Baum-Welch Algorithm, and it combines both the forward and backward algorithms. We will not go into more detail about these algorithms; additionally, there are many references to them such as [22, 34, 33, 35, 36, 37].

4.0 ARABIC DIALECT IDENTIFICATION SYSTEM

In this chapter, we present novel identification system for Arabic dialects. We first present the speech database used throughout the work. We then present a simple system utilizing GMMs and discuss its shortcomings. Later, we present a novel system based on HMMs and demonstrate the improvements achieved in dialect identification.

Figure 4.1 shows the block diagram of an Arabic dialect identification system for two dialects where the task of identification system is to classify the input speech as either Egyptian or Gulf dialect.

4.1 ARABIC DIALECT DATABASE

There is a lack of Arabic language databases in general, not only for Modern Standard Arabic (MSA), but also for other dialects of Arabic. Most databases in existence for either MSA speech recognition or Arabic Dialect identification were created solely for private research.

At present, the major standard dialect database available through the Linguistic Data Consortium (LDC)[38] includes data from the Egyptian, Levantine, Gulf, and Iraqi dialects. However, at the time we began the work only the *CALLHOME* Egyptian Arabic Speech from the LDC database was available. Therefore, we created an additional database for our work by recording TV soap operas containing both the Egyptian and Gulf dialects. Unfortunately, these recordings often contain background noises such echoes, coughs, laughter, and

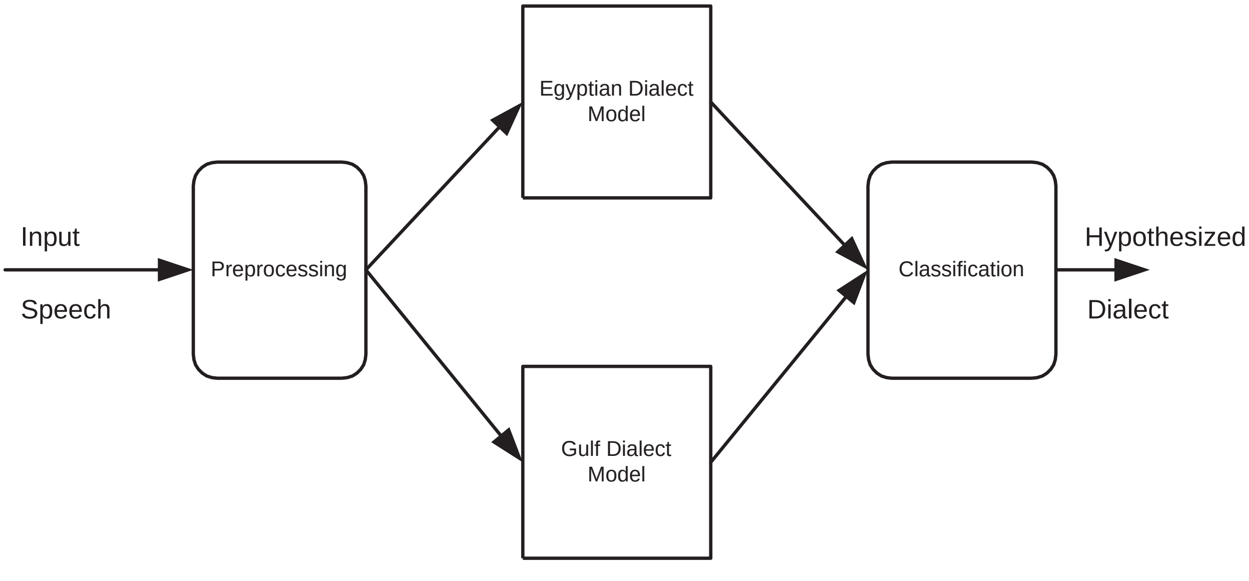

Figure 4.1: The Dialect Identification system.

background music . The overall condition of these recorded databases is poor compared to that of a standard speech database. Furthermore, this additional database contains samples from only male speakers. The speech corpus for this work consists of:

- The Egyptian Dialect: The Egyptian dialect used in this project is a combination of twenty male speakers from the CALLHOME database, and twenty male speakers from the TV recordings database. The speech of ten speakers from each database is used for training, and the speech from the other ten is used for testing. The speech for training from each speaker is one minute long.
- The Gulf Dialect: The speech used for this dialect is solely from the TV recordings database. The speech from ten male speakers is used for training, while a different set of ten speakers is used for testing.

We establish three training databases: one unbalanced and two balanced. The unbalanced training database consists of twenty Egyptian speakers (ten from the CALLHOME database and ten from recording database) and ten speakers from the Gulf dialect database. The balanced training databases, set 1 and set 2, have the same number of speakers for each dialect. For all dialects the length of speech is one minute per speaker.

For the test corpus one unbalanced database is used, consisting of twenty Egyptian speakers and ten Gulf speakers. The speakers in the test database are different from the training database, and the length of speech per speaker is 30 seconds.

It should be mentioned that no detailed or word-level labeling was done for this database; speech is labeled according to the corresponding dialect data. For instance, an Egyptian speech file is labeled with the letter "E" while a Gulf speech file is labeled with "G".

4.2 PRELIMINARILY RESULTS USING GAUSSIAN MIXTURE MODELS

At the start of the research, we conducted dialect identification experiments using Gaussian Mixture Models, GMM. The Gaussian Mixture Model, GMM, as explained in the previous chapter, Section 3.3, is defined by

$$P(\mathbf{x}|\lambda) = \sum_{i=1}^{m} w_i \; p_i(\mathbf{x}) \tag{4.1}$$

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} \, |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)' \, (\Sigma_i)^{-1} \, (\mathbf{x} - \boldsymbol{\mu}_i) \right\} \tag{4.2}$$

To start, we build three GMMs based on the three dialect databases: The CALLHOME database of Egyptian speech, TV recordings of Egyptian speech, and TV recordings of Gulf speech. The size of the sound segments in each database is one minute of speech per speaker with a total of ten different speakers in each dialect database. Gaussian models are randomly initialized and the Expectation-Maximization (EM) algorithm is used to update parameters of the GMM. The equation 4.1 is used with testing data to determine the likelihood of each

Table 4.1: Testing with part of training database for GMM (mixture size 128).

Testing with	**Training with**		
	Egyptian (CALLHOME)	**Egyptian (Recording)**	**Gulf (Recording)**
Egyptian (CALLHOME)	100%	0%	0%
Egyptian (Recording)	0%	100%	0%
Gulf (Recording)	0%	40%	60%

dialect. To do this, two test databases are used; the first consists of 30 seconds from each speaker in the training data. The second test database consists of 30 seconds of speech per speaker in the testing data, totaling ten speakers for each dialect sets, and differs from the training database.

The following tables illustrate the results of Gaussian mixture models with a mixture size of 128. The results in the tables represent the percentage of correctly identified dialects. Table 4.1 shows the results of 30 second testing segments from each speaker in the training data, while Table 4.2 shows results from the testing database.

It is easy to see that the results are degraded when the testing database is used, specifically in the case of Egyptian dialects. However, results for Gulf dialects remain identical in both instances.

If the mixture size is increased to 256, the following results can be obtained; Table 4.3 shows testing results using the whole training database instead of part it.

Table 4.4 provides the results of testing performed with the training database with a speaking time of 30 seconds per sample. As expected, dialect identification results in Table 4.4 are degraded when compared to Table 4.3, since less data is tested from the training database.

Finally, Table 4.5 displays the results for testing performed utilizing the testing database. In the final table, Table 4.5, results show degradation and the rate of identification of the

Table 4.2: Testing with testing database for GMM (mixture size 128).

Testing with	**Training with**		
	Egyptian (CALLHOME)	**Egyptian (Recording)**	**Gulf (Recording)**
Egyptian (CALLHOME)	60%	20%	20%
Egyptian (Recording)	0%	70%	30%
Gulf (Recording)	0%	40%	60%

Table 4.3: Testing with training data for GMM (mixture size 256).

Testing with	**Training with**		
	Egyptian (CALLHOME)	**Egyptian (Recording)**	**Gulf (Recording)**
Egyptian (CALLHOME)	100%	0%	0%
Egyptian (Recording)	0%	100%	0%
Gulf (Recording)	0%	20%	80%

Table 4.4: Testing with part of training data for GMM (mixture size 256).

Testing with	**Training with**		
	Egyptian (CALLHOME)	**Egyptian (Recording)**	**Gulf (Recording)**
Egyptian (CALLHOME)	100%	0%	0%
Egyptian (Recording)	0%	100%	0%
Gulf (Recording)	0%	30%	70%

Table 4.5: Testing with testing database for GMM (mixture size 256).

Testing with	**Training with**		
	Egyptian (CALLHOME)	**Egyptian (Recording)**	**Gulf (Recording)**
Egyptian (CALLHOME)	40%	60%	0%
Egyptian (Recording)	0%	100%	0%
Gulf (Recording)	0%	100%	0%

Gulf dialect drops to zero percent. In contrast, correct identification of the Egyptian dialect is high, since TV recordings from the Egyptian database dominate the identification results. Therefore, to improve the results a threshold is found and subtracted from the log probabilities of both the Egyptian and Gulf data sets. The threshold is set based on the score identification results. The ideal threshold is found at 3100, which is subtracted from P_{EMRC}; 1500 is subtracted from P_{GMRC}, where $(EMRC)$ indicates use of the Egyptian Male Recording database and $(GMRC)$ indicates use of the Gulf Male Recording database. The following tables illustrate the results after threshold reduction is performed; Table 4.6 shows testing with training data, while Table 4.7 shows test data results.

The last tables' results indicate that in the case of the Egyptian recording set there is little improvement. On the other hand, improvement is noticed in the CALLHOME and Gulf dialect sets. There are many reasons why the GMM did not work with our system, one of them being that no consideration was made for the silence segments during the modeling, possibly triggering degraded results. Also, the model for each dialect models sounds that are common to all the dialects. This produces modeling inefficiencies. The aforementioned results motivate us to further develop our proposed identification model built with silence in mind upon the Hidden Markov Model, which is discussed in the next section. Finally, Table 4.8 illustrates the results in a format that will be easy to compare to later on.

Table 4.6: Testing with training data for GMM after applying threshold (mixture size 256).

Testing (with Threshold)	**Training with**		
	Egyptian (CALLHOME)	**Egyptian (Recording)**	**Gulf (Recording)**
Egyptian (CALLHOME)	100%	0%	0%
Egyptian (Recording)	0%	100%	0%
Gulf (Recording)	0%	0%	100%

Table 4.7: Testing with testing data for GMM after applying threshold (mixture size 256).

Testing (with Threshold)	**Training with**		
	Egyptian (CALLHOME)	**Egyptian (Recording)**	**Gulf (Recording)**
Egyptian (CALLHOME)	100%	0%	0%
Egyptian (Recording)	0%	60%	40%
Gulf (Recording)	0%	40%	60%

Table 4.8: GMM results (percent correct).

Number of Mixtures	**Testing with**	
	Training Data	**Testing Data**
Mix 128	86.67	63.33
Mix 256	90	46.67
Mix 256 (with Thresholds)	100	73.33

4.3 THE HIDDEN MARKOV MODEL FOR DIALECT IDENTIFICATION

In [20], single state HMM is found to have results comparable to multistate HMMs. Yet, ergodic HMMs are shown [19] to have results comparable to Parallel Phone Recognition (PPR), which is considered to be the most popular language identification system. In the preceding chapters, we found that the Arabic dialects have similarities and differences between them; hence, there are common sounds between each dialect as well as unique sounds. For example, in the Egyptian dialect *Jiim* is pronounced as /g/ and the *Thaa* can be pronounced either *Taa* or *Saa*. We used this distinctive feature of Arabic dialects to build our system. The proposed identification system is an ergodic HMM model consisting of four states where two entry and exit states are considered to be non-emitting. Non-emitting states do not have any output probability, but they are important in joining models together for continuos speech recognition [28]. Since the identification system is based upon Ergodic HMM, there is a transition, or full connection between state 2 and 3. An ergodic model is one that, "has the property that every state can be reached from another state in a finite but aperiodic number of steps" [22].

Figure 4.2 illustrates the difference in general between the left-to-right HMM usually used in speech recognition and ergodic HMM. Figure 4.3 illustrates the proposed model for each dialect where one state (state 2) corresponds to the unique sounds of each dialect, while the other state (state 3) corresponds to common sounds across dialects.

Since there are common sounds in all Arabic dialects, we ensure that the common states share the same Gaussian mixture distribution by tying them across all dialect models. Figure 4.4 illustrates the tying process. When two states or parameters are tied, the same data is used to update their distribution (in this case the speech corresponding to the common sounds in all the dialects). The common state in each dialect model is tied across all models; therefore, the models will share the same Gaussian mixture since they all model the same sounds. A similar model was used previously by some researchers [39] for topic detection and tracking.

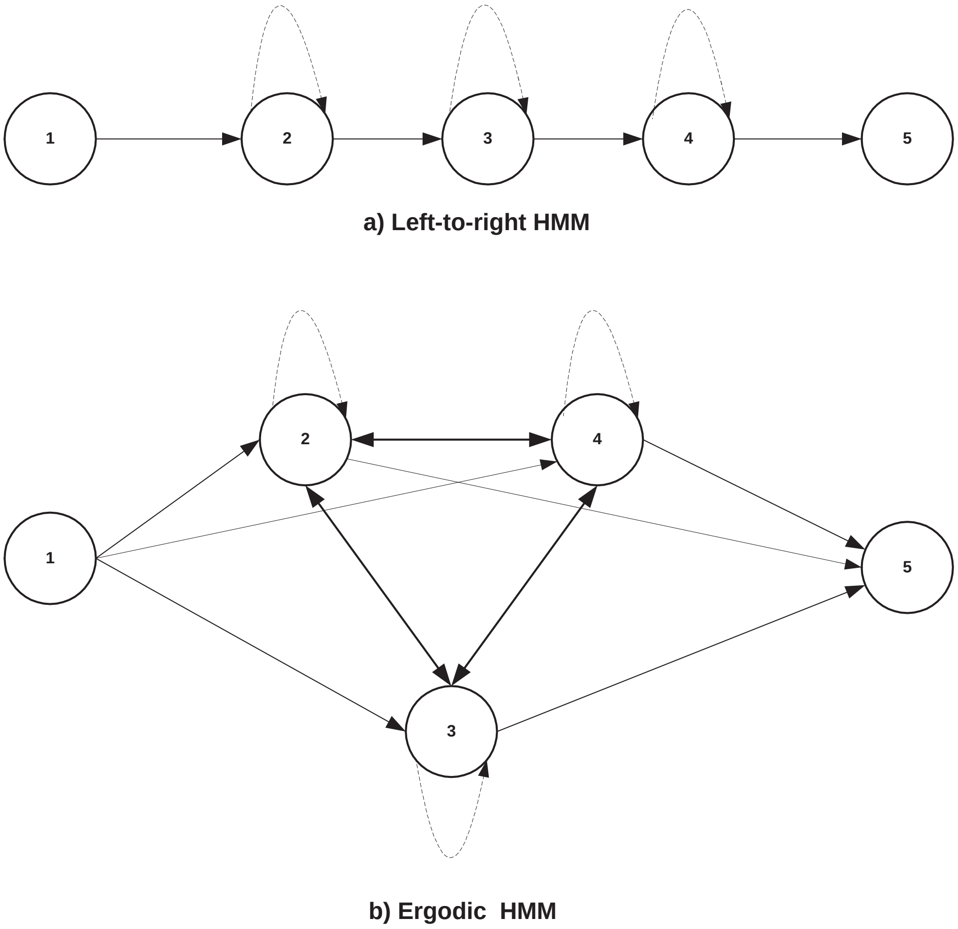

Figure 4.2: Ergodic HMM vs. left-to-right HMM.

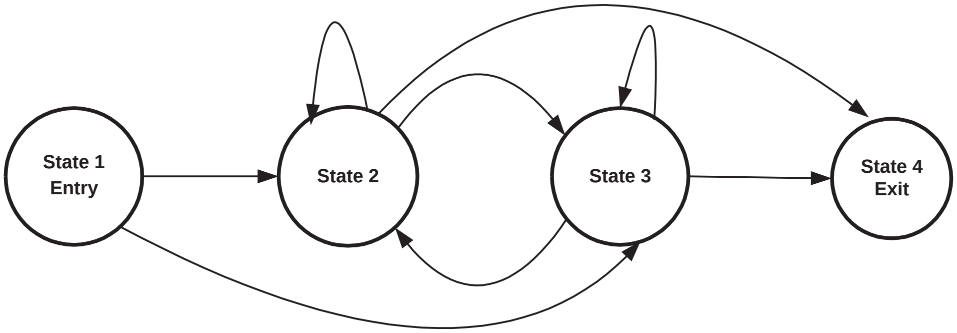

Figure 4.3: The Dialect Model.

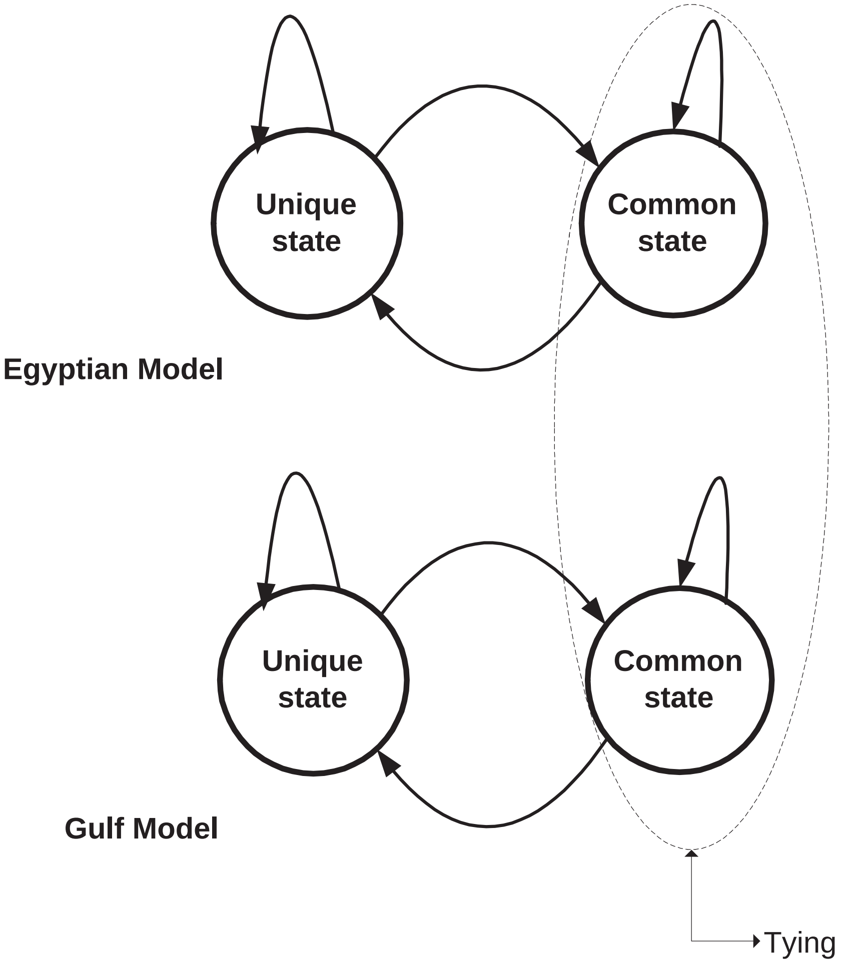

Figure 4.4: The State Tying in the Dialect Identification System.

In the re-estimation process, all data used in the estimation is combined together for a more robust estimation [28]. Moreover, in the recognition process, the computation required to decode HMM with tied parameters will often be reduced[28]. Many parameters can be tied in HMM such as the following: state distribution, mean vector, variance vector, transition matrix, and so on. In our model, state distribution is set to be tied across a selected state in each HMM.

The speech features used in this project are the Mel Frequencies Cepstral Coefficients (MFCC), coupled with cepstral mean normalization. The front-end was implemented by the **H**idden Markov Model **T**ool **K**it (HTK). HTK is a set of tools (programs) used to train the HMMs (estimate the parameters of a set of HMMs) using previously labeled speech. Then, unknown speech (Test data) is transcribed using some other HTK tools [40, 28].

In order to perform dialect identification on HTK, a language model is required. It was described in section 3.2 how recognition works with the language model. It is necessary to develop a language model capable of meeting our needs; specifically, one able to distinguish between silence and Egyptian or Gulf speech. Consequently, the language model is in the form of:

(SILENCE (E | G) SILENCE)

where *E* stands for Egyptian dialect and *G* stands for Gulf dialect.

By using this form, the language model will have only one output dialect; it begins with silence and then either becomes "E" or "G", ending in silence and avoiding a loop between dialects. This is suitable for our purposes since in the decoding process we want the recognizer to identify only one dialect for each utterance. Figure 4.5 shows the language model described above.

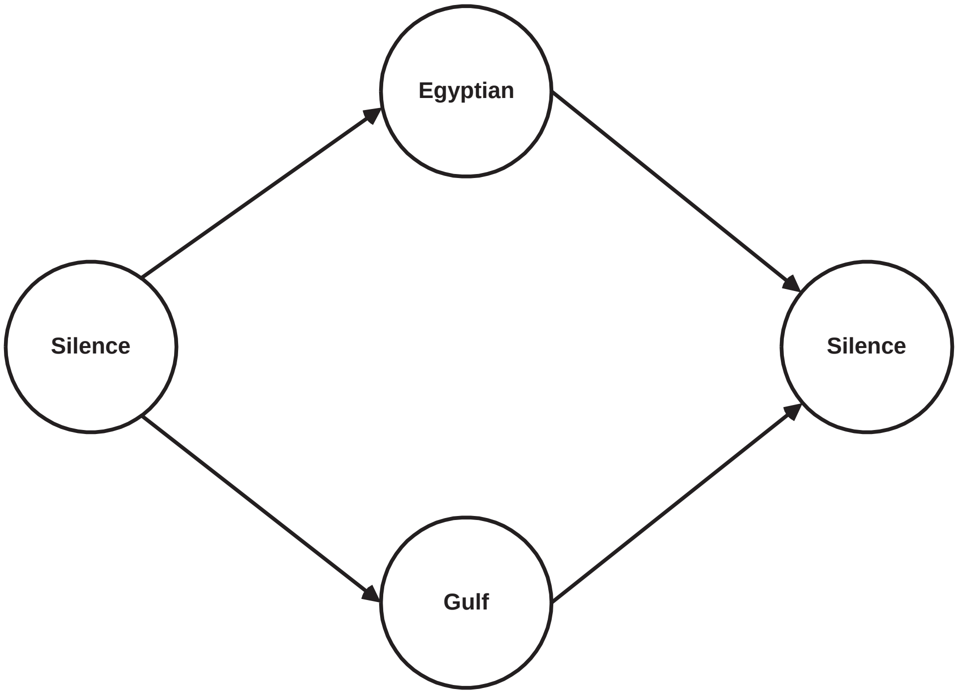

Figure 4.5: The Language Model.

4.4 PRELIMINARY RESULTS USING HIDDEN MARKOV MODELS

In this section, we build, train, and test the dialect model using the unbalanced training database and two balanced training databases. In the beginning, we use the two states model having the same mixture size. Then, we modify the dialect model either by increasing the number of states of the unique or common state or by increasing the number of mixtures in these states to reach the best model based on the identification results.

4.4.1 Training with the unbalanced database

First, we use the unbalanced training database that consists of 30 male speakers (twenty Egyptian and ten Gulf) to conduct our initial set of HMM-based dialect identification experiments. We train the dialect models so that all are configured with the same four states. The mixture size is increased from 1 to 128 by doubling the Gaussian mixture at each step. With each increment the model parameters are re-estimated a minimum of two times. The testing, or recognition process is divided into two sets of data. The first set consists of a database used for the training process, while the second database is for testing purposes, and different than the training data. In analyzing the results of the following tables we focus on the test database, however results from the training database are also considered. Table 4.9 shows the testing results with the training and testing data. In this table, the number of states are equal for the common and unique states; the total number of states in each model is four where two states are non-emitting states. As the table shows, the highest rate of identification is 76.67% using test data, with a mixture size of 64. In all following tables we examine test data, since it is the basis of evaluation for our system.

Since each state, common or unique, represents group of sounds that have different characteristics, therefore in order to model these sounds we attempt to increase the number of states in either the unique or common state to simulate the effect of different sounds with multiple states. Table 4.10 lists the results of different combinations. Two states in each model are considered unique, while the common state has only one state. Figure 4.6 illustrates the dialect model with the new states combination. The total number of states in

Table 4.9: All models have the same number of states and mixtures.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	69.50	66.67
Mix 2/2	62.41	60.00
Mix 4/4	68.09	50.00
Mix 8/8	75.18	56.67
Mix 16/16	80.14	60.00
Mix 32/32	82.62	53.33
Mix 64/64	87.59	76.67
Mix 128/128	84.40	60.00

each model is five, with two non-emitting states. The common state is tied across all models. The number of mixtures in the common state is twice that of the number of mixtures in each unique state. This keeps the number of parameters equal for the unique sounds and the common sounds. In Table 4.10, the results improve, with an optimum result of 80% in mixture 32 and mixture 64, due either to the increased number of unique states or to the increased mixture size in the common state. In addition the results for testing with the training database is improved yield to 98.94% for mixture size 128.

In Table 4.10 the unique state consists of two states, while in the following table, Table 4.11, the common states are modeled with two states; the unique state is modeled with one state. To show the new dialect model, Figure 4.7 illustrates the two tied common states and one unique state. The total number of states of each model is five, of which two are non-emitting. The two common states are tied across all models. Increasing the number of common states is more logical, since there are more common sounds than unique ones amongst all dialects. The mixture is set at 256 for the common state, while the mixture for the unique state is set to 128.

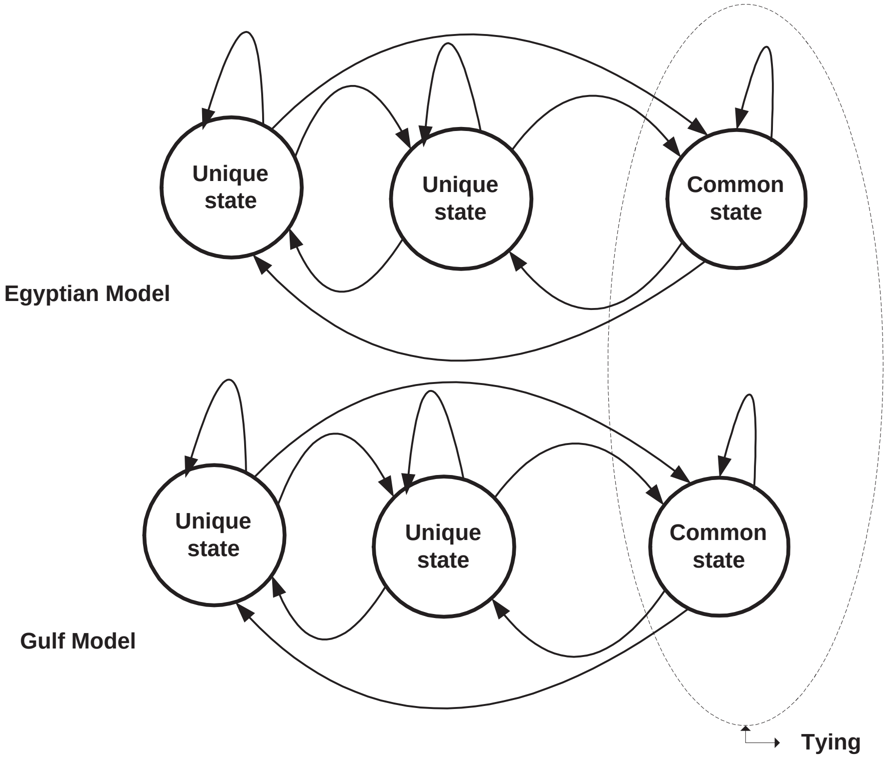

Figure 4.6: The Dialect Model of Two Unique States and One Common State.

Table 4.10: Unique state two states; common state one state.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	53.90	46.67
Mix 2/4	55.32	53.33
Mix 4/8	64.54	56.67
Mix 8/16	69.50	56.67
Mix 16/32	57.45	70.00
Mix 32/64	95.74	80.00
Mix 64/128	97.87	80.00
Mix 128/256	98.94	76.67

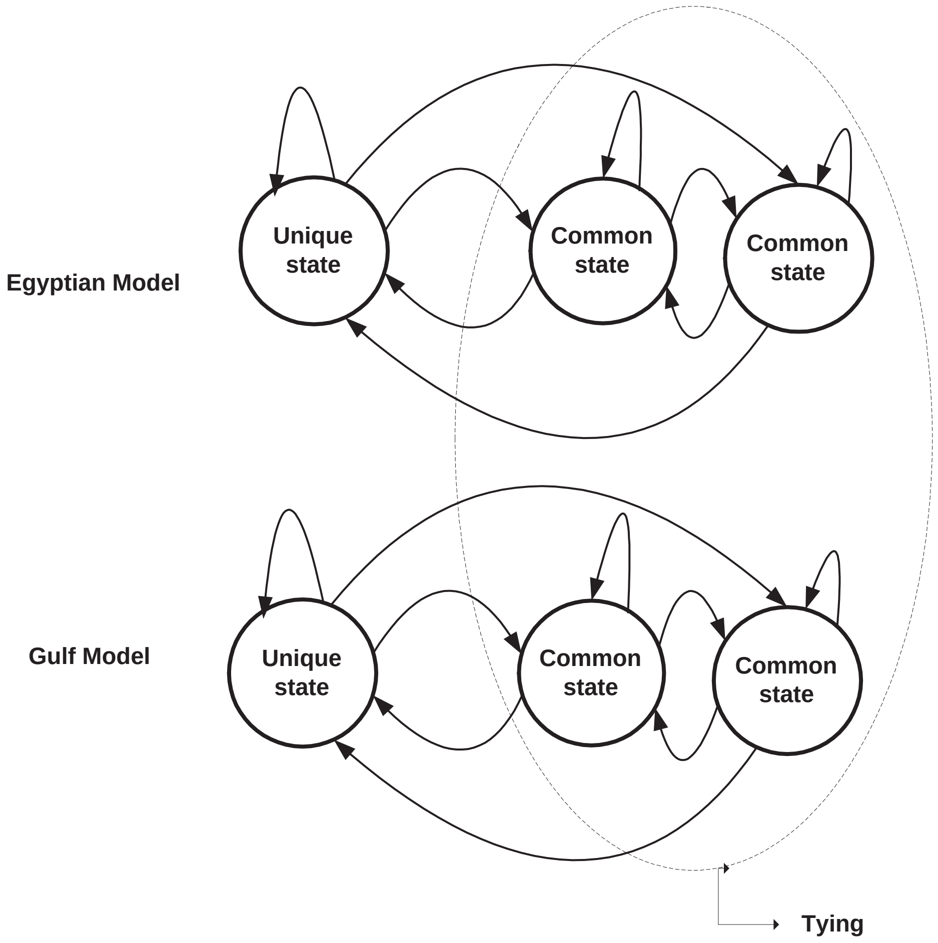

Figure 4.7: The Dialect Model of One Unique State and Two Common States.

Table 4.11: Unique state one state; common state two states.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.45	43.33
Mix 2/4	60.99	46.67
Mix 4/8	62.77	50.00
Mix 8/16	73.76	63.33
Mix 16/32	76.24	60.00
Mix 32/64	75.89	86.67
Mix 64/128	80.14	83.33
Mix 128/256	86.52	83.33

In Table 4.11 the results for testing data show improvement in comparison with the results in Table 4.10. The best identification result in Table 4.11 is 86.67% for a mixture size of 32 when testing with the test data. Results in the case of training data are degraded, and resembling the results for the test data. This behavior indicates that the models are generalizing and not just memorizing the training data.

However, if the mixture size for all states is fixed the results will change. Table 4.12 shows the results for models that consist of one unique state and two common states and an identical mixture size for all states. The highest rate of identification is 73.33% with a mixture size of 16. The layout of the model is similar to Figure 4.7.

The degradation in the results in Table 4.12 may cause one to think that increasing the mixture size is a more superior method when compared to increasing the number of states. To examine this belief further, we increase the common states to four. Figure 4.8 shows the dialect model having one unique state and four common states. The results of this modification can be examined in Table 4.13, where the mixture size is identical in all states. The highest rate of identification is 76.67% with a mixture size of 32, 64, and 256.

Table 4.12: Unique state one states; common state two states; same mixture size.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	63.83	73.33
Mix 2/2	62.41	60.00
Mix 4/4	65.50	56.67
Mix 8/8	72.34	56.67
Mix 16/16	78.72	73.33
Mix 32/32	77.66	70.00
Mix 64/64	86.52	70.00
Mix 128/128	86.88	70.00
Mix 256/256	87.23	70.00

In the next table, Table 4.14, there are eight common states and one unique state. The layout of the dialect model is similar to Figure 4.8 except that the number common states is eight. The common states are tied across all models; the mixture size is identical in all states.

In the last three tables, it is noteworthy that performance degrades as the number of common states increases. This caused us to look for increases in the mixture and then increases in the number of states, which is shown in Table 4.11 where the common states are two and also the mixture size is double. Table 4.15 shows the performance where the common state is represented by one state and so is the unique state; the number of mixtures in the common state is double that of the unique state.

The best identification result is 80% for mixture size 32/64 and 128/256. The results utilizing the training data exhibit significant improvement when compared to the results in other tables. Also, the results utilizing the test data exhibit improvement when compared to the results in Tables 4.12, 4.13, and 4.14.

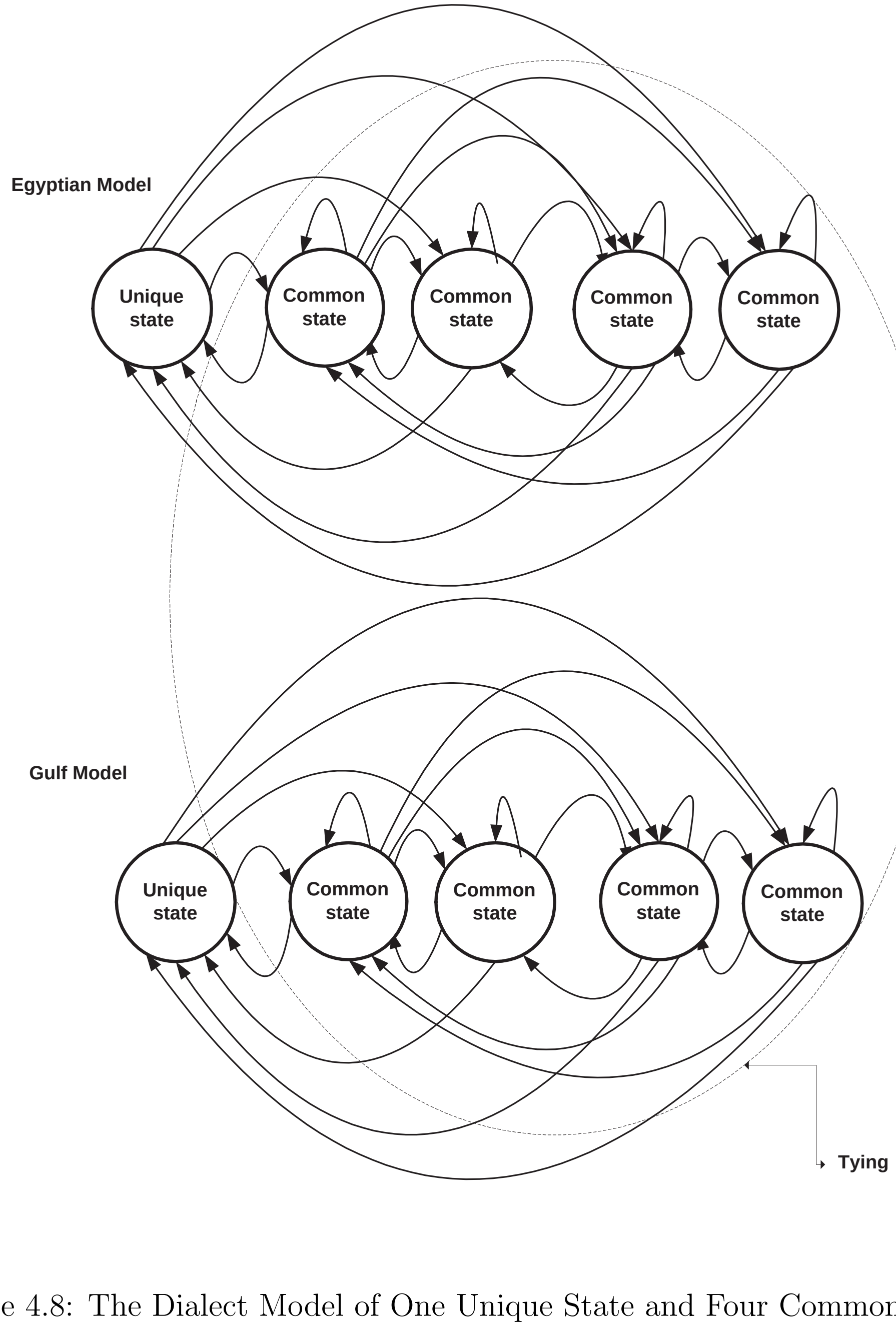

Figure 4.8: The Dialect Model of One Unique State and Four Common States.

Table 4.13: Unique state one state; common state four states; same mixture size.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	65.96	66.67
Mix 2/2	64.54	63.33
Mix 4/4	70.21	73.33
Mix 8/8	71.28	70.00
Mix 16/16	74.47	73.33
Mix 32/32	75.18	76.67
Mix 64/64	78.01	76.67
Mix 128/128	80.50	70.00
Mix 256/256	85.11	76.67

Table 4.14: Unique state one state; common state eight states; same mixture size.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	69.15	63.33
Mix 2/2	64.18	63.33
Mix 4/4	69.15	73.33
Mix 8/8	71.63	66.67
Mix 16/16	69.86	53.33
Mix 32/32	73.40	66.67
Mix 64/64	74.47	66.67
Mix 128/128	74.82	63.33
Mix 256/256	81.91	60.00

Table 4.15: Double Mixture Common State; common and unique states are one state each.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	69.50	63.33
Mix 2/4	69.86	56.67
Mix 4/8	72.70	63.33
Mix 8/16	82.27	70.00
Mix 16/32	89.72	76.67
Mix 32/64	97.52	80.00
Mix 64/128	96.81	76.67
Mix 128/256	98.23	80.00
Mix 256/512	98.58	76.67

In the following table, Table 4.16, results are shown for a case where the common state has a mixture size four times greater than the unique state and the number of states is the same in all models. The best identification result was 83.33% for the mixture sizes 16/64 and 32/128.

The next table, Table 4.17 shows the results in a case where the common state is increased in mixture size by eight times the mixture size of the unique state. The computation increases dramatically as the mixture size is increased, prompting us to stop at mixture size of 64 for the unique state, which corresponds to mixture size 512 for the common state. Also, the number of states is the same in all models.

The best identification result is 76.67% for mixtures 16/128, 32/256, and 64/512. The optimum results while utilizing test data occur when mixture of the common state is four-times greater than the mixture of the unique state, producing an identification score of 83.33%. However, an identification score of 80.00% in Table 4.15 is obtained for double mixture in

Table 4.16: Common state has four times the mixture; common and unique states are one state each.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	69.50	63.33
Mix 2/8	74.47	60.00
Mix 4/16	74.47	53.33
Mix 8/32	84.75	63.33
Mix 16/64	91.49	83.33
Mix 32/128	95.74	83.33
Mix 64/256	96.81	80.00
Mix 128/512	98.23	76.67
Mix 256/1024	98.58	76.67

Table 4.17: Common state has eight times the mixture, common and unique states are one state each.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	69.50	63.33
Mix 2/16	70.00	63.33
Mix 4/32	76.60	66.67
Mix 8/64	85.82	70.00
Mix 16/128	93.26	76.67
Mix 32/256	96.10	76.67
Mix 64/512	97.16	76.67

the common state, with a smaller computation time when compared to a common state mixture size four-times greater. Also, earlier we obtained an even better identification result of 86.67% in Table 4.11, however when compared to results in Table 4.15 while testing with training data, the results of Table 4.11 are degraded. As mentioned earlier, this degradation is likely due to a change in the number of states. Therefore, the remaining analysis will utilize the Double Mixture Common State model to achieve more balanced identification results for both the training and testing database.

4.4.2 Training with the balanced training database

In all the previous results, the training database consists of unbalanced data, meaning that the amount of Egyptian speech samples is double that of Gulf samples. The differences between a balanced and unbalanced training database were discussed previously in section 4.1. Since our database contains twenty Egyptian speakers we split it into two balanced sets. The models, Double Mixture Common State, are run under these newly balanced training databases; the results are shown in Table 4.18 using the training balanced database set number 1. Table 4.19 illustrates the results under the same conditions as above except utilizing the second set of the balanced training database.

Results show significantly improved identification with the utilization of balanced training databases; under these circumstances we obtain an identification success rate of 90.00% in Table 4.19 or 86.67% in Table 4.18. As we investigated in subsection 4.4.1, we train the models using a different number of states in the common state. Table 4.20 and Table 4.21 show the results of a common state consisting of two states and a unique state consisting of one state, while the mixture size is the same across all states.

Results for the Double Mixture Common State model in Tables 4.18 and 4.19 are significantly better than the results in the last Tables, 4.20 and 4.21, when two common states model is used and the mixture size is the same across all states. The highest identification success rate in the last two Tables, 4.20 and 4.21, is 80.00%, compared to 90.00% or 86.67% in the case of a Double Mixture Common State model. To complete this comparison, Tables 4.22 and 4.23 show the results of a case where there are four common states and only one unique state and also the mixture size is the same in all states.

Table 4.18: Double Mixture Common State (balanced data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.15	33.33
Mix 2/4	69.78	60.00
Mix 4/8	73.63	63.33
Mix 8/16	81.32	66.67
Mix 16/32	89.01	76.67
Mix 32/64	97.25	76.67
Mix 64/128	98.35	76.67
Mix 128/256	98.90	86.67
Mix 256/512	99.45	83.33

Table 4.19: Double Mixture Common State (balanced data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	53.80	66.67
Mix 2/4	59.24	63.33
Mix 4/8	73.94	63.33
Mix 8/16	83.70	66.67
Mix 16/32	93.48	76.67
Mix 32/64	94.02	76.67
Mix 64/128	96.74	86.67
Mix 128/256	98.37	83.33
Mix 256/512	98.91	90.00

Table 4.20: Two common states; One unique state (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	57.14	50.00
Mix 2/2	65.38	60.00
Mix 4/4	65.93	66.67
Mix 8/8	69.78	60.00
Mix 16/16	65.93	60.00
Mix 32/32	75.82	70.00
Mix 64/64	89.01	73.33
Mix 128/128	91.76	80.00
Mix 256/256	91.76	80.00

Table 4.21: Two common states; One unique state (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	52.17	50.00
Mix 2/2	51.63	60.00
Mix 4/4	54.89	53.33
Mix 8/8	56.52	50.00
Mix 16/16	66.30	56.67
Mix 32/32	72.28	66.67
Mix 64/64	78.80	73.33
Mix 128/128	81.52	80.00
Mix 256/256	83.70	66.67

Table 4.22: Four common states; One unique state (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	68.68	43.33
Mix 2/2	69.23	60.00
Mix 4/4	71.43	60.00
Mix 8/8	70.33	60.00
Mix 16/16	71.43	53.33
Mix 32/32	77.47	56.67
Mix 64/64	79.12	60.00
Mix 128/128	86.81	50.00
Mix 256/256	92.86	50.00

Table 4.23: Four common states; One unique state (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	49.46	56.67
Mix 2/2	51.09	56.67
Mix 4/4	50.54	50.00
Mix 8/8	55.43	43.33
Mix 16/16	70.11	66.67
Mix 32/32	75.00	60.00
Mix 64/64	84.24	50.00
Mix 128/128	79.35	63.33
Mix 256/256	84.78	66.67

From the last two Tables, 4.22 and 4.23, the best result is 66.67%. The results illustrated in these tables are definitely inferior when compared to results in the case of a Double Mixture Common State. Therefore, modeling with a Double Mixture Common State will become the basis for future experiments and improvements.

4.5 SUMMARY OF INITIAL RESULTS

In the previous sections, we showed results for two algorithms: GMM and HMM. In analyzing the results, we consider the recognition results, meaning results obtained by applying the test database. In the GMM system, the training and testing database were split into three sub databases which are: Egyptian CALLHOME, Egyptian Recording, and Gulf Recording. We have applied two mixture sizes: 128 and 256. For Mixture size 128, a 60% identification rate was obtained for the CALLHOME dialect set, 70% for the Egyptian Recording data set, and 60% for the Gulf recording data set. However, when we increase the mixture size to 256, the results degraded and the identification results were in favor of the Egyptian recording data set. For instance, the identification result of the 256 mixture was 40% for the CALLHOME data set, while in the case of the Gulf, the identification result was 0%. One of the reasons for the degradation of results may be due to a lack of modeling silence in the system and the presence of background noise in the Egyptian and Gulf recordings. However, if we apply a threshold to both the Egyptian and Gulf recordings fairer results are obtained, with identifications for both Egyptian and Gulf dialects at 60%.

Ergodic HMMs are shown to have results comparable to Parallel Phone Recognition (PPR) [19]. Therefore we present a new system with ergodic HMMs based on the features of Arabic dialects in which two states exist per model, one corresponding to the common speech in Arabic dialects and the other to the unique speech in each dialect. Two sets of experiments performed, one using unbalanced training database and the other balanced one. In the first set no preference was made for any state; all models had the same number of states and mixtures, producing an optimum result of 76.67% with a mixture size of 64. Since Arabic dialects share common sounds between them, we modeled the system with a greater number of common states, believing that there are more common sounds than unique ones. From the results, we found that increasing the mixture number is a more appropriate approach, producing a correct identification rate of 80% when the mixture of the common state was doubled. Even better results were obtainable, 86.67%, when the number of states were increased with the mixture size in the common state. However, the results of tests when the mixture was merely doubled were positive for both training and test data.

In the second set of experiments, we used a balanced training database. In the beginning mixture was doubled in the common state; the highest identification result was 90% and 86.67% for both sets of training data. If the common states were increased by a factor of two or four the results began to degrade when compared with increased mixture size in the common states. Our optimum correct identification rate of Arabic dialects was 90%.

5.0 IMPROVEMENTS TO THE ARABIC DIALECT IDENTIFICATION SYSTEM

In this chapter, we explore different alternatives to improve the system presented in the last section, Section 4.3, such as by introducing different methods of initializing the state parameters, proving that tying states model outperforms untied state models. Also, we introduce the Jackknife method to our model which provides us with more efficient use of data to train and test since we have a limited database for training and testing. Then, we utilize the Shifted Delta Cepstra features to train and test the proposed system and compare the identification results with the previous results that were obtained using the MFCC features. Finally, different combinations of speech features are used in training and testing the Arabic dialect identification system to find the best identification rate.

The outline of this chapter is as follows:

- Initialization.
- Tied dialect models vs. untied dialect models.
- Jackknifing.
- Shifted Delta Cepstra.
- Additional speech features and comparisons.

5.1 INITIALIZATION

The initialization utilized in the HMM models in section 4.4 was performed with constant vector means and variances in the states, which is considered to be a suboptimal approach

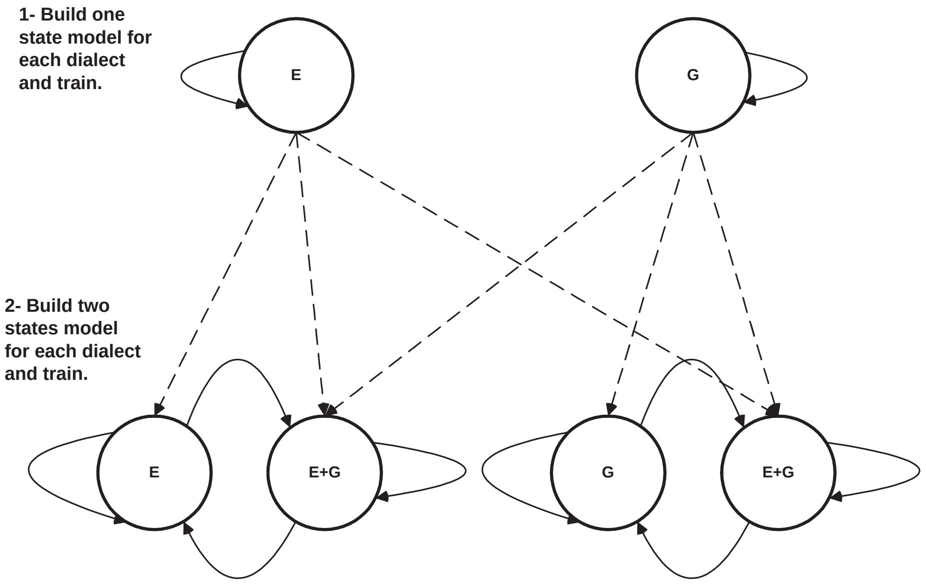
1- Build one state model for each dialect and train.
E
G
2- Build two states model for each dialect and train.
E
E+G
G
E+G

Figure 5.1: The initialization process.

[41, 42]. An alternative approach to initialization involves the use of random vector means and variances, though it is also suboptimal, since the space generated by random variables is larger than the state space [41, 42]. Because of the inadequacies in these prior methods, we apply a new approach [41, 42] of state splitting, where one selected state is split into two states, as described below:

1. A HMM model of one unique state is built for each dialect and trained.
2. Then a new HMM model of two states (unique and common) is built for each dialect, based on the previous model. The unique state in the new model is the same unique state in one unique state model, while the common state in the new model is the union of the two different unique states of the previous model.

Figure 5.1 shows the initialization process described above. This approach is more appropriate for our work since the resulting common states would include more sounds from the dialects by combining the one-state dialect models. Table 5.1 shows the results using this initialization method where the model is Double Mixture Common State, utilizing the

Table 5.1: Double Mixture Common State model initialized by Gaussian mixture components.

Number of Mixtures Unique/Common	Testing with Training Data	Testing with Testing Data
64/128	96.15	93.33
128/256	98.90	90.00
256/512	98.90	90.00

balanced training database. The best identification result on the testing data is 93.33%, which is an improvement when compared to the results in Table 4.19. Next we test the advantage of tying.

5.2 TIED DIALECT MODELS VS. UNTIED DIALECT MODELS

In the second part of this work, we compare different models in which the common states are tied and models in which there is no tying between the common states. Table 5.2 lists the results using the models without tying. The model consists of two states; the number of mixtures is the same across all models. The training data used is the un-balanced training database.

The best identification result for the un-tied model is 90% for the mixture size of 32. However, when we tie the common state across models, the results improve as shown in Table 5.3, where the model is still the two states model with a single common state, and a single unique state; the number of mixtures is the same across all models. The training data used belongs to the unbalanced training database. The best identification results on the testing data for this tied model is 93.33% for a mixture size of 32.

Table 5.2: Two state model without tying.

Number of Mixtures	Testing with	
	Training Data	Testing Data
1	69.50	63.33
2	69.50	70.00
4	73.40	66.67
8	69.50	76.67
16	86.17	83.33
32	97.52	90.00
64	98.94	83.33
128	98.94	80.00
256	99.29	76.67

Table 5.3: Two state model with tied common states.

Number of Mixtures	Testing with	
	Training Data	Testing Data
1	59.57	56.67
2	75.18	76.67
4	74.11	73.33
8	79.79	80.00
16	89.01	83.33
32	95.74	93.33
64	97.52	86.67
128	98.94	83.33
256	99.29	80.00

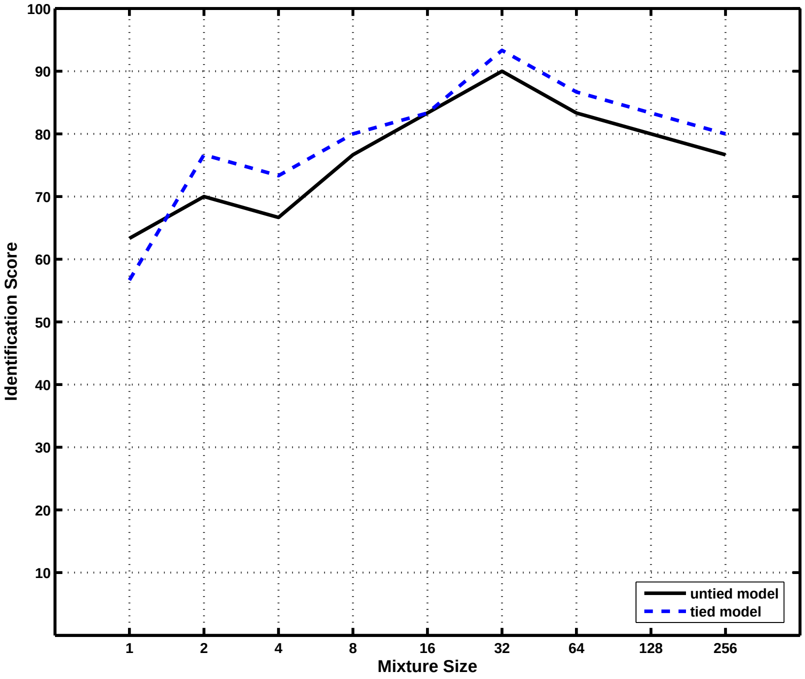

Figure 5.2: The identification system performance comparison between untied model and tied model

Figure 5.2 illustrates the performance of the dialect identification system using tied model along with the performance of the identification system using untied model. Clearly, the performance of the dialect identification system with tied model is better than the untied one from mixture size 2 and above. This improvement is expected because, by tying the common states in the tied model, these states have the same state distribution where in the case of the untied model the common state in each dialect model has its own distribution.

In the following section, a statistical method for estimating and compensating for bias [43], Jackknifing, is used.

5.3 JACKKNIFING

Jackknifing is a statistical method for obtaining an unbiased estimator [44]. Jackknifed statistics are created by systematically dropping out subsets of data one at a time and assessing the resulting variation in the studied parameters.

Assume $X_1, \cdots, X_n$ to be a random samples and let $\widehat{\Theta}$ be an estimator of the parameter θ based on the sample of size n [44]. Let X_j be removed from the samples, then the partial estimate $\widehat{\Theta}_{-j}$ is found by

$$\widehat{\Theta}_{-j} = \frac{\left(n\,\widehat{\Theta} - \Theta_j^*\right)}{(n-1)} \tag{5.1}$$

which yields

$$\Theta_j^* = n\,\widehat{\Theta} - (n-1)\,\widehat{\Theta}_{-j} \tag{5.2}$$

$$\text{for} \quad j = 1, 2, \cdots, n$$

Θ_j^* in equation 5.2 is called Pseudo-values. The average of the pseudo-values is the Jackknife of the estimate Θ

$$\widehat{\Theta}^* = \sum \frac{\widehat{\Theta}_j^*}{n} \tag{5.3}$$

Since a limited amount of training and testing data is available, the Jackknifing method provides us with more data to train and test. In this method, we combine all the databases in one database, 90% of the data is used as training data, while the remaining 10% is utilized for testing in the first iteration. In the second iteration, a different 10% of the entire data is considered as testing data; this process continues until all 10% subsets of data have been used once in the testing set.

The Double Mixture Common State model is used in this jackknife process for training and testing. From all available data, 60 speakers, speech from six speakers is used as testing and the remaining speech from 54 speakers is used as training. In this section, we report only the best identification result, the worst result, and the average result. Table 5.4 illustrates the results for the worst identification result and this corresponds to the fifth set as reported in Appendix B . Table 5.5 illustrates the results for the best identification result and this corresponds to the eighth set as shown in Appendix B. Moreover, Table 5.6, illustrates the average result across each mixture from the results of the ten sets. From the table, the best identification average result is 73.63%.

5.4 SHIFTED DELTA CEPSTRA

One of the shortcomings of the HMMs is the lack of the explicit modeling of the temporal structure of the speech features. For example, duration is poorly modeled using HMMs. The Mel-Frequency cepstral coefficients, MFCC, the most widely used features for language identification and also speech recognition, are considered to be static features [1]. To overcome this limitation, derivatives of cepstral features which capture the temporal behavior of the speech features since they measure the change in cepstral coefficients over time are added to the feature set [1, 23].

By taking the difference of the end points of the cepstral frame, delta cepstra are created as estimates of the derivatives of cepstral coefficients. Delta cepstra are considered to be dynamic features [1]. To integrate additional temporal information spanning large number

Table 5.4: The worst identification results – Jackknife results.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	18.99	100.00
Mix 2/4	52.99	77.27
Mix 4/8	58.96	68.18
Mix 8/16	91.79	20.45
Mix 16/32	94.40	18.18
Mix 32/64	97.01	13.64
Mix 64/128	98.51	11.36
Mix 128/256	99.25	4.55
Mix 256/512	99.25	2.27

Table 5.5: The best identification results – Jackknife results.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	68.63	100.00
Mix 2/4	68.30	83.33
Mix 4/8	69.61	83.33
Mix 8/16	81.37	83.33
Mix 16/32	84.37	100.00
Mix 32/64	93.46	100.00
Mix 64/128	95.42	100.00
Mix 128/256	97.39	100.00
Mix 256/512	99.02	100.00

Table 5.6: The average results – Jackknife results.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	59.70	67.32
Mix 2/4	63.40	62.68
Mix 4/8	69.86	60.94
Mix 8/16	81.07	63.84
Mix 16/32	88.54	72.14
Mix 32/64	93.28	72.35
Mix 64/128	95.91	73.63
Mix 128/256	97.91	70.61
Mix 256/512	98.97	69.22

of frames[45], the Shifted Delta Cepstra, SDC, features are used. The shifted delta cepstra are created by stacking delta cepstra features across multiple speech frames [45, 46].

The computation of the shifted detla feature is illustrated by Figure 5.3 [2]. The SDC features are specified by a set of 4 parameters: N, d, P, and k explained below [45, 46] where each feature vector contains N×k elements.

- N is the number of cepstral coefficients computed at each frame.
- P is the amount of shift between blocks.
- d is the time shift for the delta computation.
- k is the number of delta cepstra blocks used to build the Shifted Delta Cepstra feature vector.

The delta cepstra vector is found using

$$\delta_j(t) = c_j(t+d) - c_j(t-d) \tag{5.4}$$

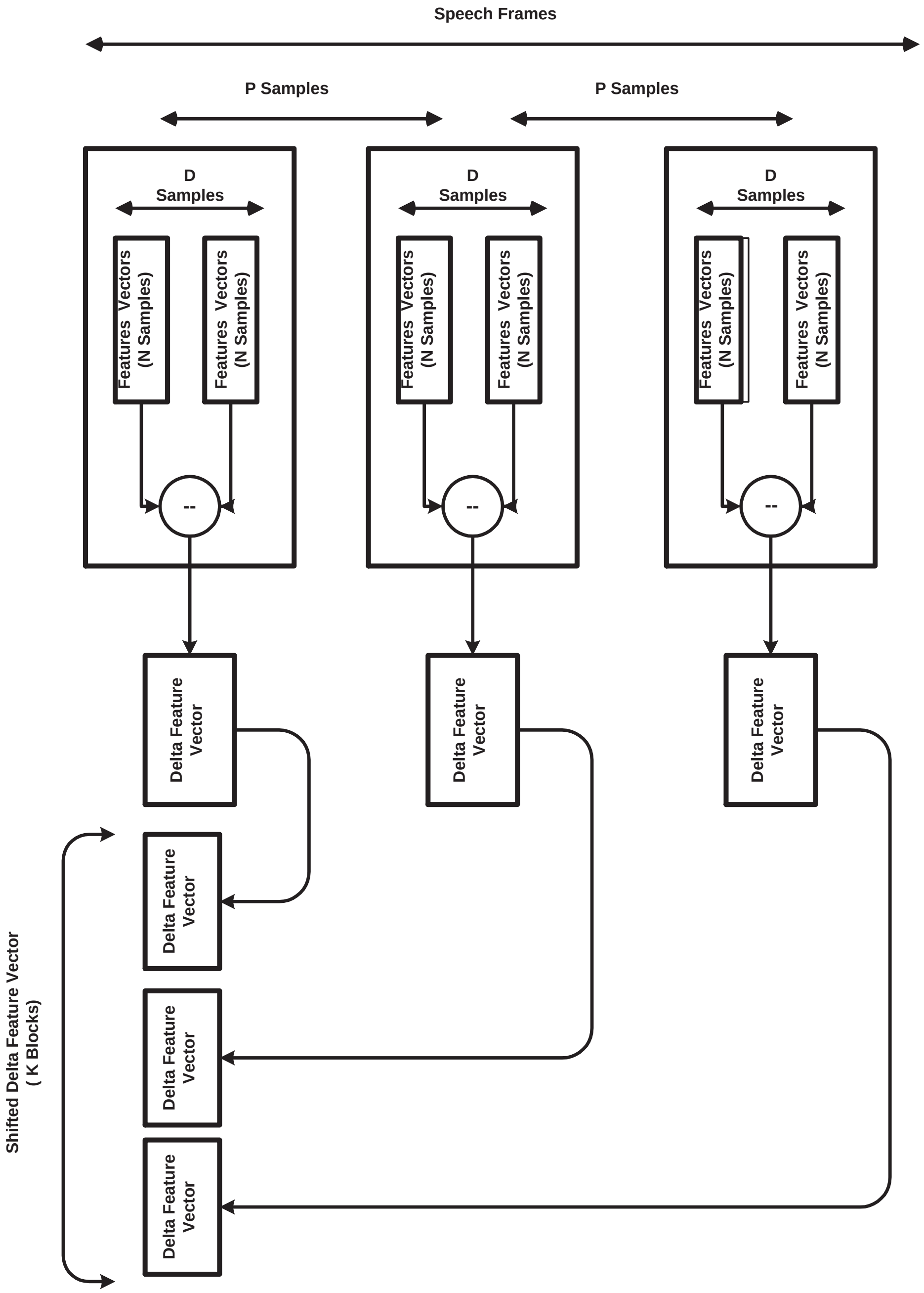

Figure 5.3: The Shifted Delta Cepstral features vector [2].

where $j = 0, \cdots, N-1$ and $c_j(r)$ is the j^{th} cepstral coefficients from the r^{th} windowed frame of speech. By concatenating k blocks of delta cepstra each shifted by P, the SDC expands the delta cepstra [45, 46]

$$sdc(t) = \delta_j(t + (i-1)P) = c_j(t + (i-1)P + d) - c_j(t + (i-1)P - d) \tag{5.5}$$

where $j = 0, \cdots, N-1$ and $i = 1, \cdots, k$

The $\delta_j(t)$ which is the delta cepstra vector is composed in the SDC vector in the following [45, 46]

$$\begin{aligned} [\delta_0(t), \delta_1(t), \cdots, \cdots, \delta_{N-1}(t) \\ \delta_0(t+P), \delta_1(t+P), \cdots, \delta_{N-1}(t+P) \\ \delta_0(t+(k-1)P), \delta_1(t), \cdots, \delta_{N-1}(t)] \end{aligned} \tag{5.6}$$

The Double Mixture Common State model was tested with MFCC features and the results were reported on Table 4.15. Using the same model, we test it with the SDC features. The parameters of SDC used in this test are 12-1-6-3, where $N = 12$, $d = 1$, $P = 6$, and $k = 3$. In the literature, a different parameter configuration, 12-1-3-3, is used. However, the previous parameter configuration leads to better dialect identification performance in our system. Using the parameters, 12-1-6-3, each frame vector has 36 features explained as the first 12 are the delta cepstra and the second 12 are the delta cepstra of the 7^{th} frame and the last 12 are the delta cepstra of the 13^{th} frame. Two feature sets are used, the first feature vectors are the MFCC features with the SDC features appended to them. The second feature vectors are the SDC features only.

As discussed in Section 4.1, three training databases are used, i.e., unbalanced database and two balanced databases to train the Double Mixture Common State model. Table 5.7

Table 5.7: MFCC + SDC features trained on unbalanced training database.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	31.21	30.00
Mix 2/4	54.26	43.33
Mix 4/8	74.47	70.00
Mix 8/16	80.50	73.33
Mix 16/32	89.36	83.33
Mix 32/64	97.16	80.00
Mix 64/128	99.29	83.33
Mix 128/256	99.29	80.00
Mix 256/512	99.29	70.00

shows the results using the unbalanced training database using the MFCC features combined with the SDC features. Each feature vector is 48 features, the first 12 features are the MFCC and the rest features, 36, are the SDC features.

The best identification in Table 5.7 is 83.33%. Using only MFCC features, the best identification as reported on Table 4.15 was 80.00%. There is an improvement in the system performance using the MFCC + SDC features over a system trained on the MFCC features.

Utilizing the balanced training databases, the best identification for train data set 1 as illustrated in Table 5.8 is 86.67%. While for the train data set 2, the best score is 83.33% as reported in Table 5.9. Comparing the system performance with a system using the MFCC features only, there is reduction in performance since the best identification result of the system trained on balanced database set 1 is 86.67% as in Table 4.18. While for training database set 2, the best performance as reported in Table 4.19 is 90.00%.

Next, the SDC features without adding to them the MFCC are utilized as the speech features. The number of features in each speech frame is 36. Table 5.10 illustrates the

Table 5.8: MFCC + SDC features trained on balanced training database – data set 1.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	48.35	30.00
Mix 2/4	61.54	33.33
Mix 4/8	68.13	53.33
Mix 8/16	83.52	70.00
Mix 16/32	90.66	83.33
Mix 32/64	98.90	83.33
Mix 64/128	99.45	86.67
Mix 128/256	99.45	86.67
Mix 256/512	99.45	83.33

Table 5.9: MFCC + SDC features trained on balanced training database – data set 2.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	50.54	33.33
Mix 2/4	52.17	46.67
Mix 4/8	58.70	50.00
Mix 8/16	72.28	66.67
Mix 16/32	91.30	66.67
Mix 32/64	96.74	76.67
Mix 64/128	98.91	83.33
Mix 128/256	99.46	73.33
Mix 256/512	99.46	73.33

Table 5.10: Double Mixture Common State model using SDC without MFCC features.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	31.56	33.33
Mix 2/4	41.84	33.33
Mix 4/8	50.35	30.00
Mix 8/16	54.96	50.00
Mix 16/32	60.64	56.67
Mix 32/64	68.44	63.33
Mix 64/128	80.85	66.67
Mix 128/256	91.13	73.33
Mix 256/512	96.10	76.67

identification results using the SDC features trained on unbalanced training database. The best identification rate using this type of features, the SDC , is 76.67% which shows a degradation when compared to using the MFCC + SDC features as in Table 5.7, or compared to using the MFCC features as in Table 4.15.

Applying the same as in the unbalanced training database case, Table 5.11 and Table 5.12, using training database set 1 and training database set 2 respectively, show the results of identification using the SDC features. The number of features per speech frame is 36. The best identification score using this type of features, SDC , which is composed of Delta features of different frames is improved when compared to using the same features but for the case of the unbalanced training database. The best identification scores are 76.67% in Table 5.11 and 80.00% in Table 5.12. There are degradations when compared to using the MFCC + SDC features as in Table 5.8 86.67% or in Table 5.9 83.33%. Even more reduction in the performance when compared using the MFCC features only where the best identification result of the system trained on balanced database set 1 is 86.67% as in Table 4.18. While for training database set 2, the best performance as reported in Table 4.19 is 90.00%.

Finally, Figure5.4 shows the comparison of the system performance using the three speech features: the MFCC, the MFCC + SDC, and the SDC. Since the results using the SDC features are not compelling, we expand our feature set to include additional time derivative and energy features.

5.5 ADDITIONAL FEATURES AND COMPARISONS

In the speech recognition and language identification field, different speech features have been used such as spectral features or prosodic ones. Since in this thesis we are concerned with the spectral features such as MFCC features, we add different features related to the MFCC such as time derivatives. The first order regression coefficients of the MFCC feature vector called Delta is included [28]. Also, the second order regression coefficients, called Delta-Delta, is included. Moreover, an energy feature will be also added to the MFCC to create an additional feature. Finally, we use the shifted delta cepstra feature to complete the comparison.

Three training database are used in the comparison, the unbalanced one and the balanced ones. More details of this training database is found in section 4.1. The model used to test in this section is the Double Mixture Common State. In this section, the features that we use are: (i) MFCC, (ii) MFCC + Energy, (iii) MFCC + Delta, (iv) MFCC + Delta + Delta-Delta, (v) MFCC + Energy + Delta, (vi) MFCC + Energy + Delta + Delta-Delta, (vii) MFCC + SDC, (viii) Delta only, (ix) Delta-Delta only, and (x) SDC.

Table 5.11: Double Mixture Common State model using SDC without MFCC features (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.15	33.33
Mix 2/4	56.04	33.33
Mix 4/8	56.04	33.33
Mix 8/16	57.14	36.67
Mix 16/32	64.84	40.00
Mix 32/64	76.37	46.67
Mix 64/128	86.26	56.67
Mix 128/256	89.56	70.00
Mix 256/512	99.45	76.67

Table 5.12: Double Mixture Common State model using SDC without MFCC features (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.74	33.33
Mix 2/4	50.00	33.33
Mix 4/8	52.17	33.33
Mix 8/16	55.43	33.33
Mix 16/32	57.61	40.00
Mix 32/64	66.85	50.00
Mix 64/128	77.72	56.67
Mix 128/256	99.46	80.00
Mix 256/512	99.46	70.00

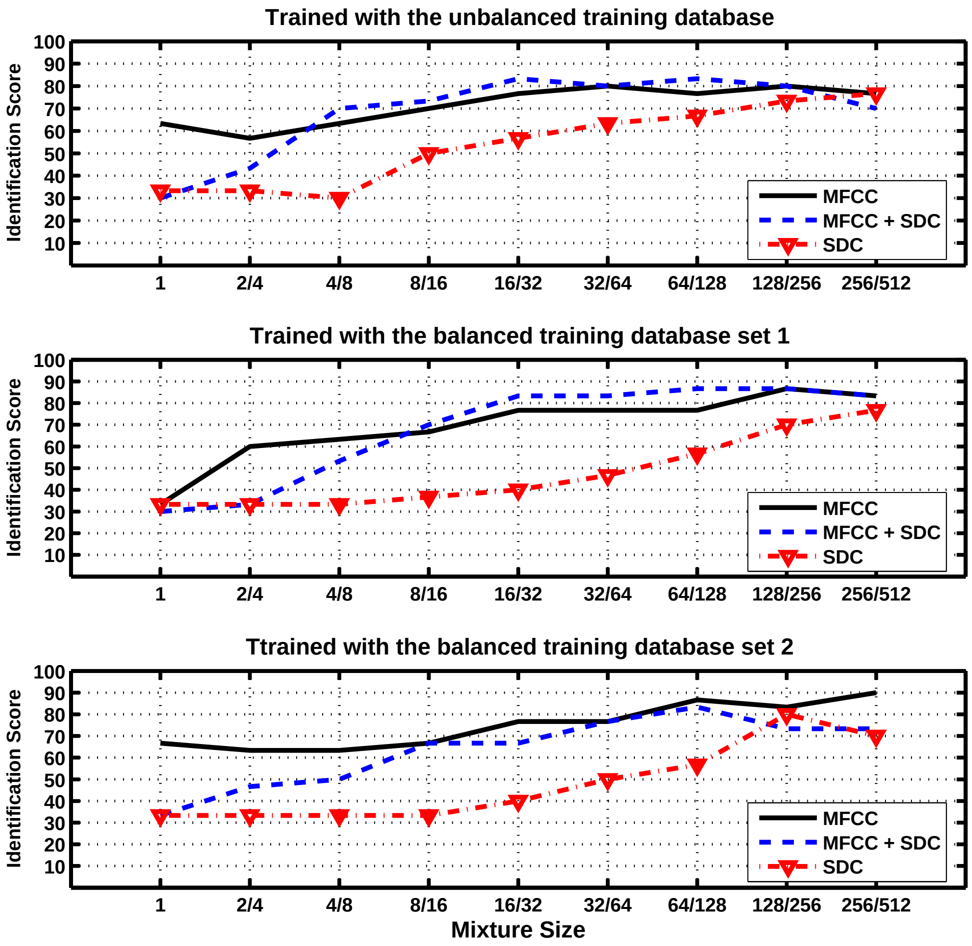

Figure 5.4: The system performance comparison for three types of features

Table 5.13: Double Mixture Common State model using MFCC features.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	69.50	63.33
Mix 2/4	69.86	56.67
Mix 4/8	72.70	63.33
Mix 8/16	82.27	70.00
Mix 16/32	89.72	76.67
Mix 32/64	97.52	80.00
Mix 64/128	96.81	76.67
Mix 128/256	98.23	80.00
Mix 256/512	98.58	76.67

5.5.1 Training with the unbalanced database

(i) MFCC:

The following table, Table 5.13, shows the results of identification using the MFCC features. The number of features per frame is 12. The best identification score utilizing these features, the MFCC, is 80.00%.

(ii) MFCC + Energy:

The energy can be computed by taking the log of the signal energy for the speech frames [28] as follow:

$$E = \log_{10} \sum_{n=1}^{N} s_n^2 \tag{5.7}$$

The log of the energy can be normalized by subtracting the maximum value of E in the speech frames [28]. This energy is appended to the MFCC features. Table 5.14 illustrates the identification results using the MFCC + Energy. The number of features per speech

Table 5.14: Double Mixture Common State model using MFCC + Energy features.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	29.79	33.33
Mix 2/4	64.54	53.33
Mix 4/8	73.40	56.67
Mix 8/16	80.50	76.67
Mix 16/32	92.20	86.67
Mix 32/64	95.39	76.67
Mix 64/128	98.94	86.67
Mix 128/256	99.29	83.33
Mix 256/512	99.29	83.33

frame is 13. The best identification score using these features, MFCC + Energy, is 86.67%. There is improvement in the identification score over the system that utilized only the MFCC features.

(iii) MFCC + Delta:

To examine more features, we add the first order regression of the MFCC, the Deltas. Table 5.15 illustrates the identification results using the MFCC + Delta features. The number of features in the speech frame is 24. The best identification score using this type of features, MFCC + Delta, is 83.33% which means that there is a reduction in the identification score when we used these features compared to the pervious table.

(iv) MFCC + Delta + Delta-Delta:

When we add Delta, and Delta-Delta, to the main features MFCC, the number of features in each speech frame is 36. Table 5.16 shows the identification results using the MFCC

Table 5.15: Double Mixture Common State model using MFCC + Delta features.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	29.79	33.33
Mix 2/4	60.28	50.00
Mix 4/8	75.89	66.67
Mix 8/16	82.62	80.00
Mix 16/32	90.78	83.33
Mix 32/64	96.45	73.33
Mix 64/128	98.23	80.00
Mix 128/256	98.94	73.33
Mix 256/512	99.28	70.00

+ Delta + Delta-Delta. The best identification rate using this type of features, MFCC + Delta + Delta-Delta, is 86.67% which is similar to the identification score found by using MFCC + Energy in Table 5.14.

(v) MFCC + Energy + Delta:

In this experiment, we add the energy and the first order regression of the MFCC, the Deltas. Table 5.17 illustrates the identification results using the MFCC + Energy + Delta features. The number of features in speech frame is 26. The best identification score using this type of features, MFCC + Energy + Delta, is 90.00% which means that there is improvement in the identification score and this is the best identification using the unbalanced database.

(vi) MFCC + Energy + Delta + Delta-Delta:

Table 5.18 illustrates the identification results using the MFCC + Energy + Delta + Delta-Delta. The number of features in each speech frame is 39. The best identification rate

Table 5.16: Double Mixture Common State model using MFCC + Delta + Delta-Delta features.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	29.79	33.33
Mix 2/4	58.51	56.67
Mix 4/8	61.70	63.33
Mix 8/16	85.11	73.33
Mix 16/32	87.59	80.00
Mix 32/64	92.20	80.00
Mix 64/128	97.87	76.67
Mix 128/256	98.94	86.67
Mix 256/512	98.94	73.33

Table 5.17: Double Mixture Common State model using MFCC + Energy + Delta features.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	29.79	33.33
Mix 2/4	57.09	46.67
Mix 4/8	72.70	70.00
Mix 8/16	80.14	83.33
Mix 16/32	88.30	90.00
Mix 32/64	95.04	90.00
Mix 64/128	98.23	90.00
Mix 128/256	99.58	90.00
Mix 256/512	98.94	86.67

using this type of features, MFCC + Energy + Delta + Delta-Delta , is 86.67%. There is a reduction in the rate of the best identification compared to the case the MFCC + Energy + Delta features as in Table 5.17.

(vii) MFCC + SDC:

Using the MFCC + SDC features as explained earlier in section 5.4, the number of features in speech frame is 48. As reported in Table 5.7, the best identification rate using this type of feature, MFCC + SDC, is 83.33%.

(viii) Delta only:

To examine the idea of using different features than the MFCC features, Delta only features are used as the main features. Table 5.19 illustrates the identification results using only the Delta features. The number of features in each speech frame is 12. The best identification rate using this type of features, Delta only, is 66.67%. There is a degradation compared to using the MFCC features as in Table 5.13 and a greater degradation when compared to the MFCC + Energy + Delta features listed in Table 5.17.

(ix) Delta-Delta only:

Furthermore Table 5.20 illustrates the identification results using only the Delta-Delta features. The number of features in each speech frame is 12. The best identification score using this type of features, Delta-Delta only, is 76.67% which shows a degradation when compared to using the MFCC features as in Table 5.13, or more reduction compared to the MFCC + Energy + Delta features as in Table 5.17.

(x) SDC :

Finally, since the SDC features were created by appending Delta cepstra features of different frames to the static MFCC features, we examine using only these Delta features without adding the MFCC features as explained earlier in section 5.4. Table 5.10 illustrates the identification results using the SDC features. The number of features in each speech frame is 36. The best identification rate using this type of features, the SDC, is 76.67% which shows a degradation when compared to using the MFCC features as in Table 5.13, or the MFCC + Energy + Delta features as in Table 5.17.

Table 5.18: Double Mixture Common State model using MFCC + Energy + Delta + Delta-Delta features.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	33.69	36.67
Mix 2/4	52.84	56.67
Mix 4/8	65.96	76.67
Mix 8/16	83.33	86.67
Mix 16/32	94.68	86.67
Mix 32/64	97.16	83.33
Mix 64/128	98.23	86.67
Mix 128/256	99.29	80.00
Mix 256/512	99.29	76.67

Table 5.19: Double Mixture Common State model using Delta only features.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	29.79	33.33
Mix 2/4	41.13	36.67
Mix 4/8	38.30	33.33
Mix 8/16	42.91	36.67
Mix 16/32	62.77	43.33
Mix 32/64	80.50	50.00
Mix 64/128	92.91	50.00
Mix 128/256	98.58	60.00
Mix 256/512	99.29	66.67

Table 5.20: Double Mixture Common State model using Delta-Delta only features.

Number of Mixtures	**Testing with**	
Unique/Common	**Training Data**	**Testing Data**
Mix 1/1	29.79	33.33
Mix 2/4	51.00	33.33
Mix 4/8	43.62	33.33
Mix 8/16	59.22	43.33
Mix 16/32	74.47	60.00
Mix 32/64	74.47	56.67
Mix 64/128	79.08	66.67
Mix 128/256	87.59	76.67
Mix 256/512	93.97	80.00

In summary, the following Figure 5.5 summarizes all the results using the features used in this section. Also, Figure 5.6 shows the comparison of best identification scores using different features for the unbalanced training database. The performance of dialect identification system improves when more features per speech are used compared to the performance of using only the MFCC features as the main features.

5.5.2 Training with the balanced database

The description of the balanced database is found in section 4.1. The balanced database consists of two sets and it will be used to train and test the model based on different speech features. The speech features used in this section are the same ones used in the previous section. The model used in this section is the Double Mixture Common State.

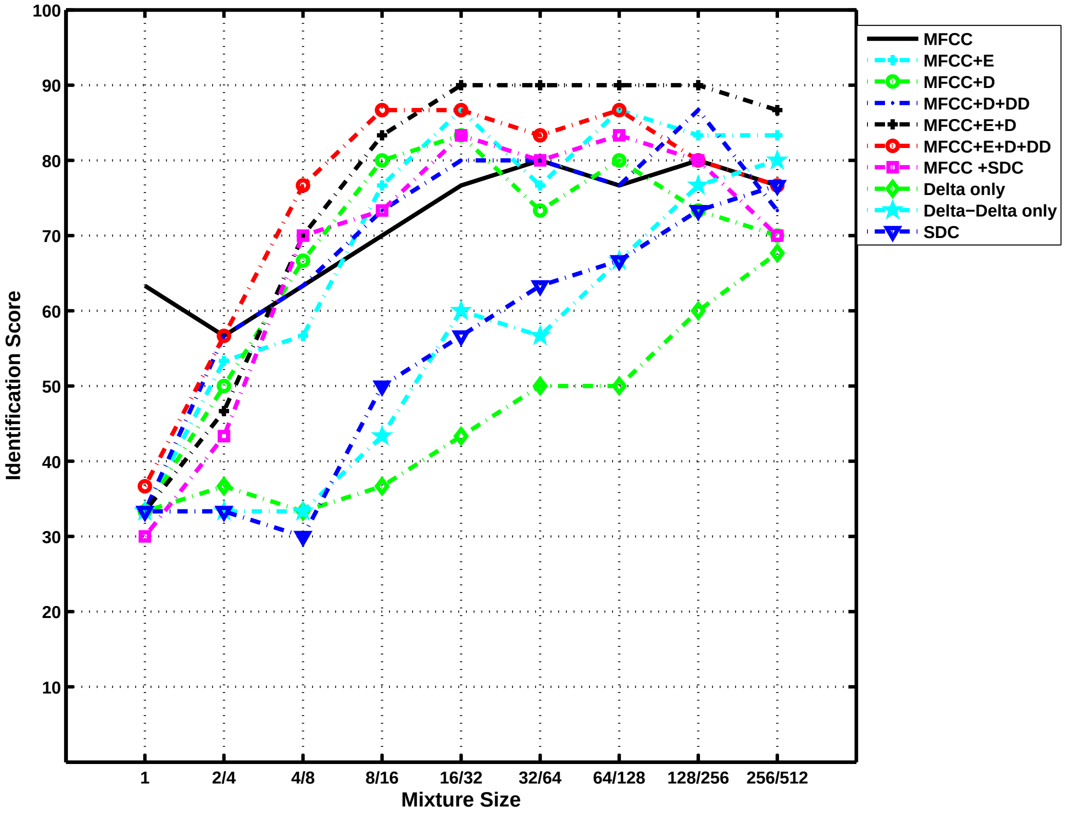

Figure 5.5: Summary results for unbalanced data.

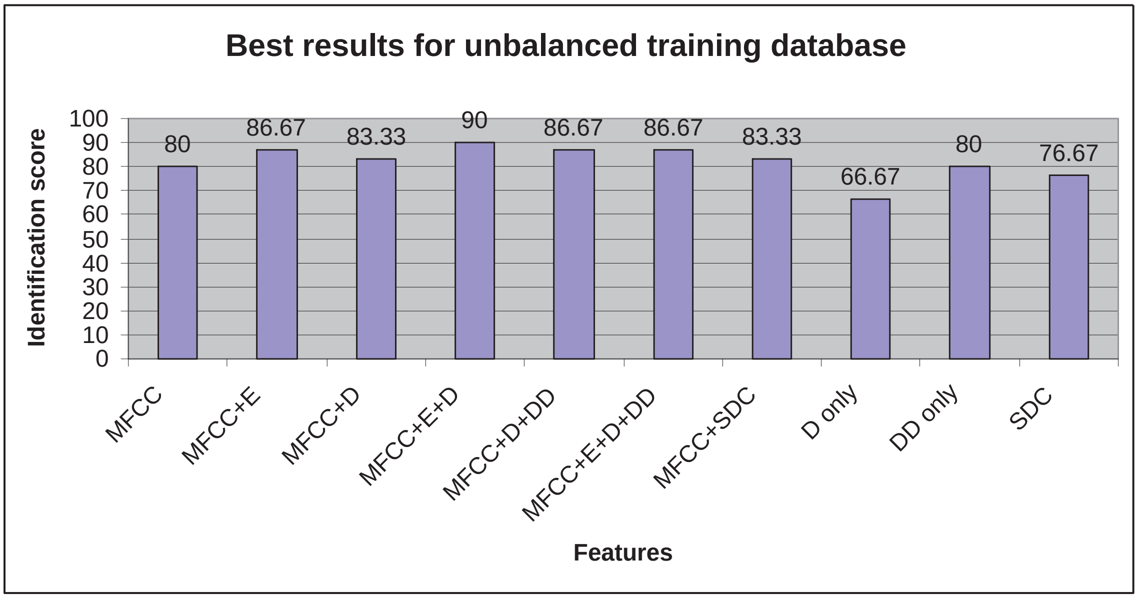

Figure 5.6: Summary of best results for un-balanced training database.

(i) MFCC:

The following tables, Table 5.21 and Table 5.22, using training database set 1 and training database set 2 respectively, show the results of identification using the MFCC features. The number of features per speech frame is 12. The best identification scores using this type of features, MFCC, are 86.67% in Table 5.21 and 90.00% in Table 5.22. As mentioned earlier in last the chapter, when using balanced databases the best identification rate improves.

(ii) MFCC + Energy:

The energy as explained in the previous section is appended to the MFCC features and the following tables, Table 5.23 and Table 5.24, using training database set 1 and training database set 2 respectively, show the results of identification using the MFCC + Energy features. The number of features per speech frame is 13. The best identification rates using this type of features, MFCC + Energy, are 83.33% in Table 5.23 and 83.33% in Table 5.24. The best identification rate shows degradation when it compared to 86.67% as in Table 5.14 for using the unbalanced training database.

Table 5.21: Double Mixture Common State model using MFCC features (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.15	33.33
Mix 2/4	69.78	60.00
Mix 4/8	73.33	63.33
Mix 8/16	81.32	66.67
Mix 16/32	89.01	76.67
Mix 32/64	97.25	76.67
Mix 64/128	98.35	76.67
Mix 128/256	98.90	86.67
Mix 256/512	99.45	83.33

Table 5.22: Double Mixture Common State model using MFCC features (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	53.80	66.67
Mix 2/4	59.24	63.33
Mix 4/8	73.94	63.33
Mix 8/16	83.70	66.67
Mix 16/32	93.48	76.67
Mix 32/64	94.02	76.67
Mix 64/128	96.74	86.67
Mix 128/256	98.37	83.33
Mix 256/512	98.91	90.00

(iii) MFCC + Delta:

Using training database set 1 and training database set 2 respectively, Table 5.25 and Table 5.26, show the results of identification using the MFCC + Delta features. The number of features per speech frame is 24. The best identification scores using this type of features, MFCC + Delta, are 86.67% in Table 5.25 and 90.00% in Table 5.26. The best identification score shows improvement when it compared to 83.33% as in Table 5.15 for using the unbalanced training database.

(iv) MFCC + Delta + Delta-Delta:

The subsequent tables, Table 5.27 and Table 5.28, using training database set 1 and training database set 2 respectively, show the results of identification using the MFCC + Delta + Delta-Delta features. The number of features per speech frame is 36. The best identification rates using this type of features, MFCC + Delta + Delta Delta, are 96.67% in Table 5.27 and 86.67% in Table 5.28. The best identification rate shows a huge improvement when compared to 86.67% as in Table 5.16 using the unbalanced training database. By far, this is the best identification rate we achieve through this work.

(v) MFCC + Energy + Delta:

Using the MFCC + Energy + Delta features, Table 5.29 and Table 5.30 show the results of dialect identification. The number of features per speech frame is 26. The best identification score using this type of features, MFCC + Energy + Delta , are 90.00% in Table 5.29 and 93.33% in Table 5.30. The best identification score shows improvement when it is compared to 90.00% as in Table 5.17 for the unbalanced training database.

(vi) MFCC + Energy + Delta + Delta-Delta:

Using the MFCC + Energy + Delta + Delta-Delta features, Table 5.31 and Table 5.32 show the results of dialect identification. The number of features per speech frame is 39. The best identification rate using this type of features, MFCC + Energy + Delta + Delta-Delta, are 86.67% in Table 5.31 and 93.33% in Table 5.32. The best identification rate shows improvement when it compared to 86.67% as in Table 5.18 for the unbalanced training database.

Table 5.23: Double Mixture Common State model using MFCC + Energy features (dataset1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.70	33.33
Mix 2/4	70.33	60.00
Mix 4/8	73.63	50.00
Mix 8/16	90.66	76.67
Mix 16/32	93.41	70.00
Mix 32/64	97.25	76.67
Mix 64/128	97.80	83.33
Mix 128/256	97.80	80.00
Mix 256/512	99.45	83.33

Table 5.24: Double Mixture Common State model using MFCC + Energy features (dataset2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	58.70	70.00
Mix 2/4	63.04	56.67
Mix 4/8	74.46	63.33
Mix 8/16	78.26	63.33
Mix 16/32	94.02	76.67
Mix 32/64	96.20	83.33
Mix 64/128	98.37	80.00
Mix 128/256	98.91	80.00
Mix 256/512	99.46	73.33

Table 5.25: Double Mixture Common State model using MFCC + Delta features (dataset1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	47.25	30.00
Mix 2/4	58.24	53.33
Mix 4/8	69.23	50.00
Mix 8/16	84.62	73.33
Mix 16/32	90.66	86.67
Mix 32/64	96.70	86.67
Mix 64/128	97.80	83.33
Mix 128/256	99.45	76.67
Mix 256/512	99.45	76.67

Table 5.26: Double Mixture Common State model using MFCC + Delta features (dataset2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	45.65	33.33
Mix 2/4	71.20	66.67
Mix 4/8	72.28	56.67
Mix 8/16	76.67	70.00
Mix 16/32	90.22	73.33
Mix 32/64	97.83	83.33
Mix 64/128	98.91	80.00
Mix 128/256	99.46	80.00
Mix 256/512	99.46	90.00

Table 5.27: Double Mixture Common State model using MFCC + Delta + Delta-Delta features (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	49.45	36.67
Mix 2/4	58.79	40.00
Mix 4/8	55.49	33.33
Mix 8/16	58.79	40.00
Mix 16/32	90.66	80.00
Mix 32/64	97.80	86.67
Mix 64/128	98.35	96.67
Mix 128/256	99.45	93.33
Mix 256/512	99.45	90.00

Table 5.28: Double Mixture Common State model using MFCC + Delta + Delta-Delta features (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.20	33.33
Mix 2/4	69.02	50.00
Mix 4/8	63.04	53.33
Mix 8/16	78.26	56.67
Mix 16/32	89.13	80.00
Mix 32/64	94.57	80.00
Mix 64/128	97.28	83.33
Mix 128/256	98.91	86.67
Mix 256/512	99.46	83.33

Table 5.29: Double Mixture Common State model using MFCC + Energy + Delta features (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	52.20	33.33
Mix 2/4	57.14	60.00
Mix 4/8	68.68	56.67
Mix 8/16	78.57	60.00
Mix 16/32	93.96	80.00
Mix 32/64	98.35	80.00
Mix 64/128	99.45	90.00
Mix 128/256	99.45	90.00
Mix 256/512	99.45	86.67

Table 5.30: Double Mixture Common State model using MFCC + Energy + Delta features (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	45.65	33.33
Mix 2/4	65.22	56.67
Mix 4/8	70.11	70.00
Mix 8/16	79.35	70.00
Mix 16/32	93.48	93.33
Mix 32/64	98.91	86.67
Mix 64/128	99.46	80.00
Mix 128/256	99.46	83.33
Mix 256/512	99.46	83.33

(vii) MFCC + SDC:

The MFCC + SDC features were explained earlier in section 5.4, the number of features in each speech frame is 48. As reported in Section 5.4, the best identification scores using this type of features,MFCC + SDC, are 86.67% in Table 5.8 and 83.33% in Table 5.9. The best identification score shows improvement when compared to 83.33% as in Table 5.7 using the unbalanced training database.

(viii) Delta only:

Utilizing the Delta only features, Table 5.33 and Table 5.34, using training database set 1 and training database set 2 respectively, show the results of identification. The number of features per speech frame is 12. The best identification rates using this type of features, Delta only, are 56.67% in Table 5.33 and 50.00% in Table 5.34. Using these Delta only features, the identification system has poor performance in both training databases, the unbalanced and balanced, compared to the score of the system trained on the MFCC features.

(ix) Delta-Delta only:

The following tables, Table 5.35 and Table 5.36, show the results of identification using only the Delta-Delta features. The number of features per speech frame is 12. The best identification score using this type of features, Delta-Delta only, is 40.00% in Table 5.35 and 40.00% in Table 5.36. Comparing the results with those obtained by the system trained on the unbalanced training database, the identification results show degradation since the best result in that case is 76.67% as in Table 5.20.

(x) SDC:

Applying the same as in the unbalanced training database case, the SDC features, were explained earlier in section 5.4, are used as the speech features for the identification system. Table 5.11 and Table 5.12, using training database set 1 and training database set 2 respectively, show the results of identification using the SDC features. The number of features per speech frame is 36. The best identification score using this type of features, SDC, which is

Table 5.31: Double Mixture Common State model using MFCC + Energy + Delta + Delta-Delta features (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	54.95	33.33
Mix 2/4	55.49	63.33
Mix 4/8	69.78	53.33
Mix 8/16	85.16	76.67
Mix 16/32	92.31	80.00
Mix 32/64	97.25	83.33
Mix 64/128	99.45	86.67
Mix 128/256	99.45	86.67
Mix 256/512	99.45	83.33

Table 5.32: Double Mixture Common State model using MFCC + Energy + Delta + Delta-Delta features (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	47.83	36.67
Mix 2/4	67.93	66.67
Mix 4/8	68.48	76.67
Mix 8/16	69.57	80.00
Mix 16/32	93.48	90.00
Mix 32/64	97.28	90.00
Mix 64/128	98.37	90.00
Mix 128/256	99.46	93.33
Mix 256/512	99.46	90.00

Table 5.33: Double Mixture Common State model using Delta only features (data-set1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.15	33.33
Mix 2/4	57.14	30.00
Mix 4/8	58.24	30.00
Mix 8/16	58.79	23.33
Mix 16/32	65.93	43.33
Mix 32/64	83.52	46.67
Mix 64/128	95.60	56.67
Mix 128/256	99.45	56.67
Mix 256/512	99.45	50.00

Table 5.34: Double Mixture Common State model using Delta only features (data-set2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	45.65	33.33
Mix 2/4	52.72	36.67
Mix 4/8	50.00	36.67
Mix 8/16	63.04	23.33
Mix 16/32	69.57	43.33
Mix 32/64	81.52	50.00
Mix 64/128	95.11	50.00
Mix 128/256	98.91	50.00
Mix 256/512	99.46	50.00

Table 5.35: Double Mixture Common State model using Delta-Delta only features (dataset1).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	46.15	33.33
Mix 2/4	48.35	33.33
Mix 4/8	51.65	36.67
Mix 8/16	52.20	36.67
Mix 16/32	64.84	30.00
Mix 32/64	68.68	30.00
Mix 64/128	73.08	36.67
Mix 128/256	84.07	40.00
Mix 256/512	95.05	40.00

Table 5.36: Double Mixture Common State model using Delta-Delta only features (dataset2).

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	45.65	33.33
Mix 2/4	64.13	36.67
Mix 4/8	57.07	36.67
Mix 8/16	50.00	33.33
Mix 16/32	64.13	36.67
Mix 32/64	66.30	33.33
Mix 64/128	75.00	40.00
Mix 128/256	88.04	40.00
Mix 256/512	95.65	36.67

composed of Delta features of different frames is improved when compared to using the same features but for the case of the unbalanced training database. The best identification scores are 76.67% in Table 5.11 and 80.00% in Table 5.12. The identification rate for the case of using the unbalanced training database is 76.67% as in Table 5.10.

In summary, the following Figures, 5.7 and 5.8 summarize all the test results for both balanced training databases. Also, Figure 5.9 and Figure 5.10 show the comparison of best identification scores using different features for both balanced training databases.

5.6 SUMMARY OF IMPROVEMENT METHODS

In this chapter, we used different methods to improve the proposed system, the two states, common and unique, ergodic Hidden Markov Model described in Section 4.3. These methods include initialization, jackknifing, shifted delta cepstra. Lastly, we concluded the chapter with a comparison with utilizing the proposed system with different features such as MFCC, MFCC + Energy, MFCC + Delta, MFCC + Delta + Delta-Delta, MFCC + Energy + Delta + Delta-Delta, MFCC + SDC, Delta only, Delta-Delta only, and SDC.

In the initialization experiment, we used algorithm described in [41, 42]. The algorithm includes state splitting into two states by building a one-state model for each dialect then using that state model to build our proposed two state model, unique and common. The unique state of specific dialect will be the one state model for that dialect while the common state will be the union of the one state models of each dialect yielding to double mixture and combination of these states. This is done for each mixture size, but we are reporting only from mixture sizes 64/128 and up. Then the built model is trained at least ten iterations. The best identification as reported in Table 5.1 is 93.33% an improvement over a Double Mixture Common State model without initialization of at least 3.33%.

In the second part of this chapter, we proved that the tying process in the proposed model leads to better results over a system without tying. At certain mixture sizes, the

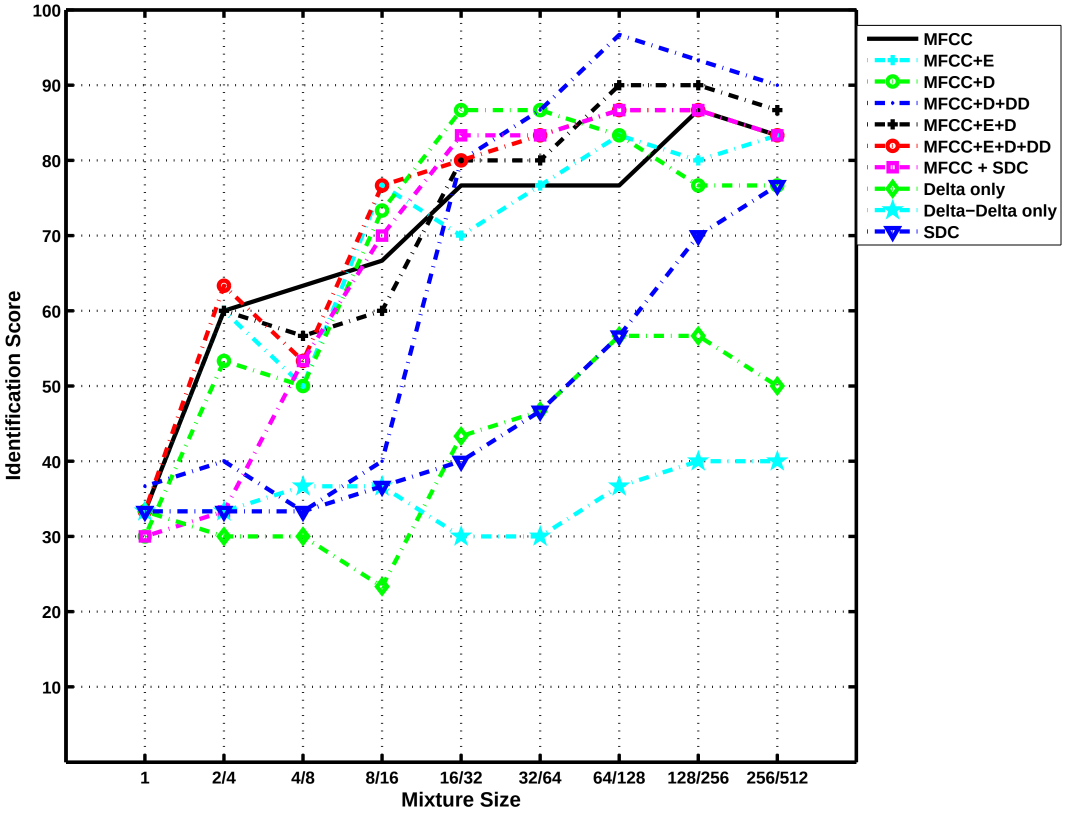

Figure 5.7: Summary results for balanced data set-1.

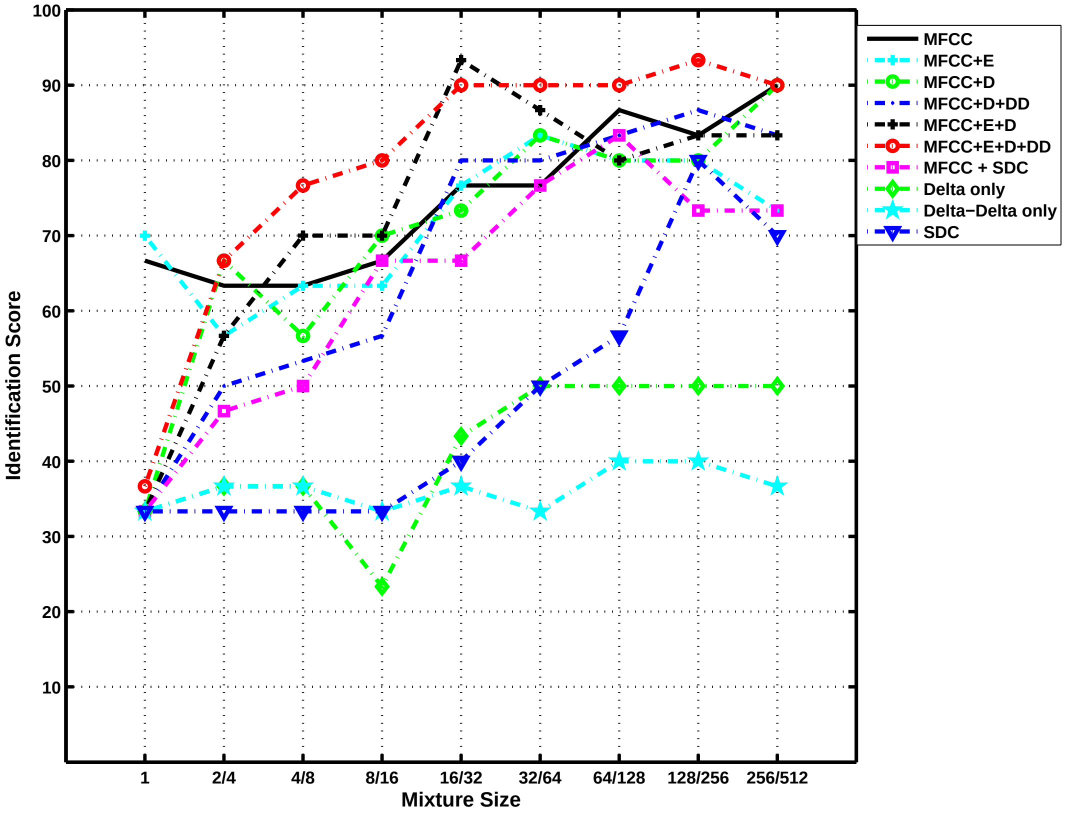

Figure 5.8: Summary results for balanced data set-2.

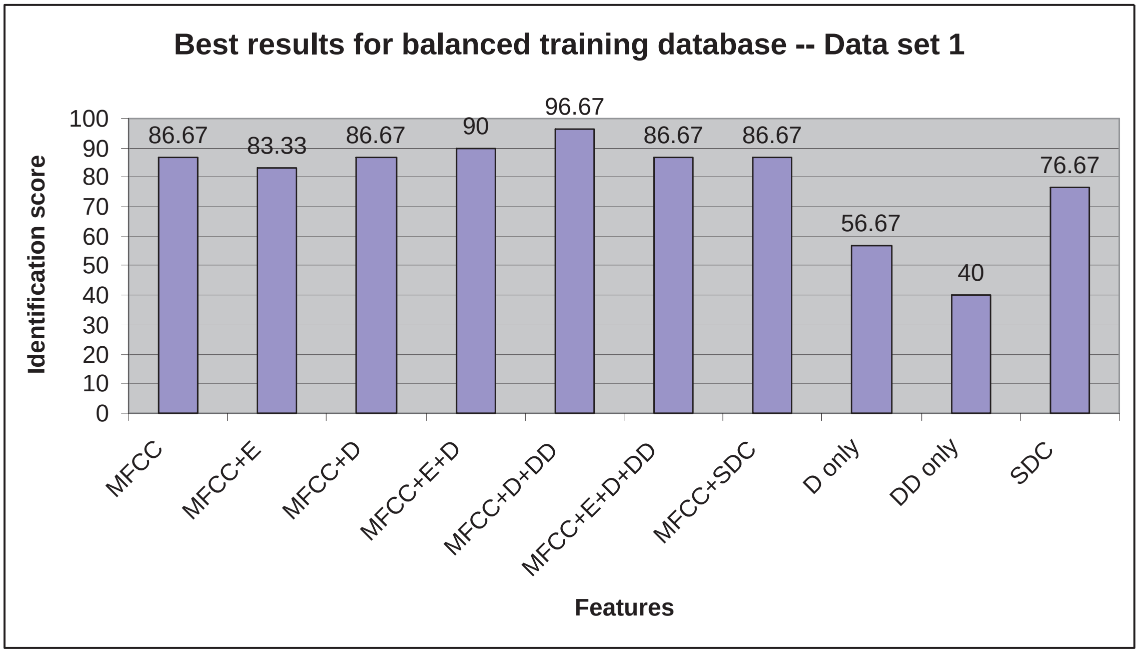

Figure 5.9: Summary of best results for balanced data set-1.

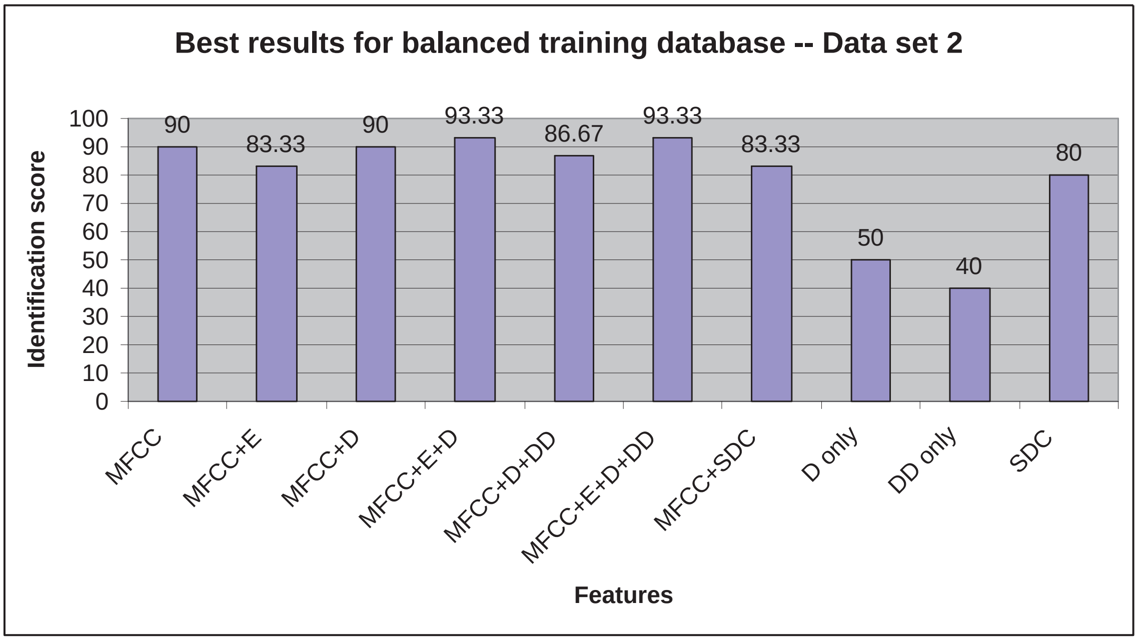

Figure 5.10: Summary of best results for balanced data set-2.

common state of the un-tied model is tied and then both the untied model and the tied model are re-estimated at least 10 times. It should be noted that the mixture size is the same across all states, unique and common. The best identification of tied model is 93.33% while the best identification of untied model is 90.00%.

We considered Jackknifing as a method for obtaining an unbiased estimator. Moreover, Jackknifing algorithm provides with greater data to train and test by dividing the whole database into sets. Set 1 is considered to the training data and consists of 90% of the database while the second set is the testing data which is the remaining 10%. In each iteration, the test set covers another 10% until it passes through the entire database. The best average identification scores for all 10 iterations is 73.63%.

Shifted Delta Cepstral, SDC, have proven to be better features when used with GMM for language identification [46, 45]. The SDC features are created by stacking delta cepstra features across multiple speech frames. The parameters of SDC used in the training and testing are 12-1-3-6, where $N = 12$, $d = 1$, $P = 6$, and $k = 3$. Two training databases were used, balanced and unbalanced. The model used is the Double Mixture Common State. Two feature vectors are used the MFCC+ SDC and the SDC. The best identification score for a system trained on the unbalanced database is 83.33% using the MFCC + SDC; while using the SDC features the best identification score is 76.67%. For the balanced training database set 1, the best performance shown in Table 5.8 is 86.67% while for database set 2 the best score is 83.33% using the MFCC + SDC features. However, in the case utilizing the SDC features, the best identification score for balanced training database set 1 is 76.67% while for the balanced training database set 2 the best identification score is 80.00%. In comparison with a system utilizing only the MFCC features, there is an improvement in system trained on the unbalanced database, however in the case of the balanced training database, there is a degradation. Also, adding the MFCC features to the SDC improve the performance of the system.

In the last section, different features are used in training and testing the system. Three training databases, unbalanced and balanced ones, are used. The best identification rate reported in this thesis is 96.67% in the case of using the MFCC + Delta + Delta-Delta features when the balanced training database set 1 is used as training data. Also, we achieve an identification score of 93.33% using the features MFCC + Energy + Delta + Delta-Delta when the balanced training database set 2 is used as training data. In the case of using the unbalanced training database, the best score found is 90.00% for the case of using the MFCC+ Energy + Delta as speech features. The dialect identification system performs better with more features per speech frame.

5.7 DISCUSSION

The uniqueness about the presented Arabic dialect identification system that it has the capability to model the common sounds in all Arabic dialects in addition modeling the unique sounds in each dialect. Although the dialect identification system was not tested with dialects from different language, we believe that the presented dialect identification system will work for any language similar to the Arabic language as long as the langauge has dialects that share common sounds and that have unique sounds.

Starting with the structure of two states, common and unique, ergodic Hidden Markov Model for the Arabic dialect identification system, we found by experiment that by doubling the mixture size in the common state compared to the unique state, the performance of the system improved. Also, we found that by using different training database that the best performance of the dialect identification system occur with the balanced training database. Due to the lack of a standard database for all Arabic dialect, we concentrate our work on two dialects: Egyptian and Gulf. It should be mentioned that the audio quality of most our training database is poor compared the standard speech database; making these training databases more realistic to real life applications.

Time derivative features such as Delta and Delta-Delta have been found to improve the performance of speech recognition system [22]. The time derivatives of the MFCC are usually appended to the feature vector in order to capture the dynamics of speech. The energy within the frame is also an important feature that can be appended to the MFCC features. Many speech recognition systems use the the logarithm of energy to compresses the dynamic range of values which is the similar to the characteristic of human hearing system [21]. Adding time derivatives and log energy to the MFCC features demonstrate a clear improvements to the performance of our Arabic dialect identification system. This implies that using more features per speech frame, 39, as opposed to using only the MFCC features, 12, enhances the performance of the system.

The initialization process is important for ergodic HMM performance [42]. Initializing with constant values which is the simplest method proved to be unsatisfactory approach. Applying the initialization method provided in [41, 42], the dialect identification system performed better.

Reaching an identification rate of 96.67%, the Arabic dialect identification system has performed superior compared to the initial results of the same system trained on the MFCC features only. The outstanding performance of the system has utilized the same training database, balanced training database set 1, with the only difference that the speech features used are the MFCC + Delta + Delta-Delta features per speech frame.

6.0 CONCLUSION AND FUTURE WORK

This thesis presented a system that automatically identifies the Arabic dialects. The difficult task of properly identifying various Arabic dialects was examined. Since the Arabic language has many different dialects, they must be identified before Automatic Speech Recognition can take place. Due to the limited availability of Arabic speech databases, it was necessary to create new data for two dialects: Egyptian and Gulf, in addition to using the CALLHOME databases created by LDC.

A new model has been presented in this work based upon the features of Arabic dialects; namely, a model that recognizes the similarities and differences between each dialect. The model utilized in this work was the ergodic Hidden Markov Model, which is comprised of two states: one representing common Arabic sounds, and another representing unique sounds characteristic of certain dialects. Common states were tied across all models since they share similar sounds. Results were provided for different types of training data, both balanced and unbalanced, and according to different system structures, such as Double Mixture Common State, and by increasing the number of common states. Moreover, improvements to the dialect identification system have been made. A new initialization process is used and yields better system performance of 93.33%. Features of different types were used in training and testing the double mixture common state model and found that the system with the MFCC + Delta + Delta-Delta features performed best reaching an identification score of 96.67% when the balanced training database, rather than an unbalanced one, is utilized.

6.1 FUTURE WORK

The task of improving Arabic dialect identification in our proposed system remains a constant area for further research. Thus, in this section we provide some ideas that can be considered for future examination; these ideas are related to concepts already presented in this thesis, yet take steps to improve performance and reduce computational costs.

- The lack of a well-established large speech database of all Arabic dialects and modern standard Arabic in general is an ongoing challenge that needs to be rectified.
- This work should be extended to include additional Arabic dialects. Future investigation into the structure of our dialect identification system should be done.
- The main emphasis in this thesis was on the proposed system not speech features; further investigation should be done to find if temporal differences in the dialect features of the dialects can be captured with small frames in the front end.
- The system presented in this study should be extended by applying a discriminative training algorithm, such as Maximum Mutual Information (MMI) used to update HMM model parameters instead of Maximum Likelihood Estimation (MLE).
- Further studies should identify an optimal length for the dialect identification of speech segments. In other words, how long should a speech segment be to contain enough unique sounds for dialect identification.

APPENDIX A

ARABIC LANGUAGE STRUCTURE

Arabic is one of the world's oldest Semitic languages, and it is the fifth most widely used. Arabic is the primary language of countries: Iraq, Syria, Jordan, Lebanon, Palestine, Saudi Arabia, Bahrain, Kuwait, Qatar, United Arab Emirates, Oman, Yemen, Egypt, Sudan, Comoros, Djibouti, Somalia, Libya, Tunisia, Algeria, Morocco, and Mauritania. Arabic is the language of communication in official discourse, teaching, religious activities, and in literature. Additionally, Arabic shares its alphabet with several other languages, such as Farsi, Urdu, and Malay [6, 47, 9].

This chapter provides background information about the Arabic language, its alphabet, phonemes, and the differences between it and English.

A.1 ARABIC LANGUAGE ALPHABET

Every language is typically partitioned into two broad categories: vowels and consonants. Vowels are produced without obstructing air flow through the vocal tract, while consonants involve significant obstruction, creating a nosier sound with weaker amplitude [9, 48]. The Arabic language consists of of 28 letters, which are shown in Table A1.

The letter *Alif* (æ) has no sound of its own, but it is used to support *Hamzah* (?) and lengthen a preceding vowel; it is also used at the end of a third person plural verb in the

Table A1: Arabic Phonemes.

Arabic Letters	Key Word	Approximate Pronunciation	English Equivalent	IPA Symbol
أ	أب	Hamzah		ʔ
ا		Alif	à	æ
ب	بدر	Baa	B	b
ت	تمر	*Taa	T	t
ث	ثواب	*Thaa	Th as in “three”	θ
ج	جمل	Jemm	J	ʤ / g / ʒ
ح	حمل	Haa	-	ħ
خ	خال	Khaa	-	χ
د	درس	*Daal	D	d
ذ	ذاكر	*Thaal	Th as in “then”	ð
ر	رزق	*Raa	R	r
ز	زمن	*Zaa	Z	z
س	سار	*Seen	S	s
ش	شمس	*Sheen	Sh	ʃ
ص	صبح	*Saad	-	s̲
ض	ضرب	Dhaad	-	d̲
ط	طب	*Tah	-	t̲
ظ	ظهر	*Dhaa	-	ð̲
ع	عبن	*Ein	-	ʕ
غ	غرب	Ghein	-	ɣ
ف	فال	Faa	F	f
ق	قمر	Qaaf	-	q
ك	كتب	Kaaf	K	k
ل	لوح	*Laam	L	l
م	ماء	Meem	M	m
ن	نور	*Noon	N	n
هـ	هلال	Haa	H	h
و	وقف	Wow	W	w / u
ي	يقف	Yaa	Y	j / i

past tense. Arabic is written from right to left and letters take different forms depending on their position in a word; some letters are similar to others except for diacritical points placed above or beneath them [49, 3].

Arab linguists classify Arabic letters into two categories: sun and moon. Sun letters are indicated by an asterisk as shown in Table A1. When the sun letters are preceded by the prefix *Alif-Laam* in nouns, the *Laam* consonant is not pronounced [49, 9].

The Arabic language has six different vowels, three short and three long. The short vowels are *fatha* (a), short *kasrah* (i), and short *dammah* (u). No special letters are assigned to the short vowels, however special marks and diacritical notations above and beneath the consonants are used. The three long vowels are durational allophones of the above short vowels, as in mad, meet, and soon and correspond to long *fatha*, long *kasrah*, and long *dammah* respectively. Consonants can be also un-vowelised (not followed by a vowel); in this case a diacritic *sakoon* is placed above the consonant. Vowels and their IPA (International Phonetic Alphabet) equivalents are shown in the Table A2 [9, 3].

A.2 ARABIC PHONEMES

Speech sound units, known as phonemes, are classified as either consonants or vowels. This classification is based upon articulator, acoustic, and contextual information [9, 3].

Standard Arabic has 34 phonemes, six of which are vowels and 28 that are consonants. Arabic phonemes can be shown as in table A3:

Voice stops are produced by releasing pressure built up after the oral tract has been completely closed. Unvoiced stops are produced in a similar fashion, except during the closure process the vocal cords do not vibrate. Fricatives in both the voiced and unvoiced stops are concentrated at high frequencies where the vocal tract is partly opened at the point of articulation and turbulent noise is created [9]. The emphatic sounds found in stops and fricatives depend upon a buccal phenomenon that occurs due to the lowering of the tongue as its base moves backward. Nasal sounds, which are voiced, are produced by a shrunk vocal

Table A2: Arabic Vowels.

Arabic Notation	Vowel	IPA Symbol	Key word
ـَ	**Short Fatha**	**a.**	**Duck**
ـِ	**Short Kasrah**	**i.**	**In**
ـُ	**Short Dammah**	**u.**	**Look**
ـا	**Long Fatha**	**a.:**	**Dad**
ـي	**Long Kasrah**	**i.:**	**Meet**
ـو	**Long Dammah**	**u.:**	**Soon**

Table A3: Arabic Consonant Phonemes.

Arabic Consonant Phonemes	

Stops	**Voiced** { b/ب ; d / د }
	Unvoiced { t/ت ; k/ك ; q/ق ;ʔ /ء }
	Emphatic { t̲/ ط ; d̲/ ض }

Fricatives	**Voiced** { ð / ذ ; z/ز ; ʤ/ج ; ʕ/ع ; ɣ/غ }
	Unvoiced { Ɵ/ ث ; f/ف ; s / س ; ʃ/ش ; X/خ ; h̲/ح ; h/هـ }
	Emphatic { ð̲/ظ ; s̲/ص }

Nasals	{ n / ن ; m/م }

Others	**Lateral** { l/ل }
	Trill { r/ر }

Vowels	{ a/ـَـ ; u/ـُـ ; i/ـِـ }
Semi-Vowels	{ j / ي ; w / و }

tract and excited glottal [9]. Vowels are voiced phonemes that are produced by semi-periodic pulses of air caused by vibrations of the vocal cords, which subsequently stimulate the vocal tract [3, 47].

A.3 DIFFERENCES BETWEEN ARABIC AND ENGLISH

Arabic and English differ significantly; some of the major differences are outlined below [3, 47]:

- Arabic has ten pharyngeal and emphatic sounds while English has none.
- Arabic has a voiced consonant *Baa* (b), while English has both voiced and unvoiced versions, resulting in a distinction between (b) and (p).
- Arabic has an unvoiced consonant *Faa* (f) while English has both voiced and unvoiced versions, resulting in the (v) and (f) sounds.
- Arabic has only six fundamental vowels while American English has twelve.
- There are some phonemes found in English that are not Arabic, such as the /g/ in game or the /c/ in sing.
- Arabic has the unique phoneme known as *Dhaad* ($\underline{\text{d}}$), which cannot be found in English or any other language.

APPENDIX B

THE RESULTS OF JACKKNIFING METHOD

In this appendix, we are going to report all the results for the Jackknife method. In section 5.3, Jackknife method was introduced and the whole database was spilt to two sets one for training and the other one for testing. The following tables show the results of training and testing with Double mixture common state model. From all available data, 60 speakers, speech from six speakers was used as testing and the remaining speech from 54 speakers was used as training.

Tables, B1 up to B10, illustrate the results for the first set up to the tenth set. Last table, Table B11 shows the average results for all the ten sets.

Table B1: Results for the first set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	61.02	98.28
Mix 2/4	58.27	86.21
Mix 4/8	64.17	94.83
Mix 8/16	78.35	89.66
Mix 16/32	83.07	96.55
Mix 32/64	92.91	96.55
Mix 64/128	96.46	96.55
Mix 128/256	98.82	98.28
Mix 256/512	98.82	98.28

Table B2: Results for the second set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	63.10	95.00
Mix 2/4	61.51	66.67
Mix 4/8	72.62	83.33
Mix 8/16	85.32	73.33
Mix 16/32	92.06	78.33
Mix 32/64	95.63	76.67
Mix 64/128	97.62	76.67
Mix 128/256	98.02	78.33
Mix 256/512	99.21	80.00

Table B3: Results for the third set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	62.30	98.33
Mix 2/4	61.90	68.33
Mix 4/8	71.83	46.67
Mix 8/16	77.38	55.00
Mix 16/32	86.90	61.67
Mix 32/64	89.29	43.33
Mix 64/128	91.67	38.33
Mix 128/256	95.63	31.67
Mix 256/512	98.41	45.00

Table B4: Results for the fourth set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	77.78	31.67
Mix 2/4	77.38	28.33
Mix 4/8	77.38	25.00
Mix 8/16	83.73	33.33
Mix 16/32	88.49	50.00
Mix 32/64	90.08	43.33
Mix 64/128	94.44	46.67
Mix 128/256	98.81	43.33
Mix 256/512	99.21	33.33

Table B5: Results for the fifth set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	18.99	100.00
Mix 2/4	52.99	77.27
Mix 4/8	58.96	68.18
Mix 8/16	91.79	20.45
Mix 16/32	94.40	18.18
Mix 32/64	97.01	13.64
Mix 64/128	98.51	11.36
Mix 128/256	99.25	4.55
Mix 256/512	99.25	2.27

Table B6: Results for the sixth set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	34.97	0.00
Mix 2/4	65.36	50.00
Mix 4/8	74.84	66.67
Mix 8/16	78.43	100.00
Mix 16/32	85.62	100.00
Mix 32/64	90.85	100.00
Mix 64/128	96.41	100.00
Mix 128/256	98.04	100.00
Mix 256/512	99.35	100.00

Table B7: Results for the seventh set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	69.97	100.00
Mix 2/4	55.23	50.00
Mix 4/8	66.99	66.67
Mix 8/16	83.33	66.67
Mix 16/32	89.54	66.67
Mix 32/64	92.48	100.00
Mix 64/128	95.75	100.00
Mix 128/256	97.06	100.00
Mix 256/512	98.37	100.00

Table B8: Results for the eight set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	68.63	100.00
Mix 2/4	68.30	83.33
Mix 4/8	69.61	83.33
Mix 8/16	81.37	83.33
Mix 16/32	84.37	100.00
Mix 32/64	93.46	100.00
Mix 64/128	95.42	100.00
Mix 128/256	97.39	100.00
Mix 256/512	99.02	100.00

Table B9: Results for the ninth set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	69.61	33.33
Mix 2/4	70.92	66.67
Mix 4/8	68.63	33.33
Mix 8/16	77.12	50.00
Mix 16/32	90.52	100.00
Mix 32/64	95.42	100.00
Mix 64/128	96.41	100.00
Mix 128/256	98.04	83.33
Mix 256/512	99.02	66.67

Table B10: Results for the tenth set.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	70.92	16.67
Mix 2/4	62.09	50.00
Mix 4/8	73.53	50.00
Mix 8/16	79.74	66.67
Mix 16/32	89.87	50.00
Mix 32/64	95.10	50.00
Mix 64/128	96.73	66.67
Mix 128/256	98.04	66.67
Mix 256/512	99.02	66.67

Table B11: The average results – Jackknife results.

Number of Mixtures Unique/Common	**Testing with**	
	Training Data	**Testing Data**
Mix 1/1	59.70	67.32
Mix 2/4	63.40	62.68
Mix 4/8	69.86	60.94
Mix 8/16	81.07	63.84
Mix 16/32	88.54	72.14
Mix 32/64	93.28	72.35
Mix 64/128	95.91	73.63
Mix 128/256	97.91	70.61
Mix 256/512	98.97	69.22

BIBLIOGRAPHY

[1] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, January 2001.

[2] Jonathan Lareau, "Application of shifted delta cepstral features for gmm language identification," M.S. thesis, Rochester Insititute of Technology, 2006.

[3] Yousef Al-Otaibi, *Pharyngeal and Emphatic sounds in Arabic Speech Recognition*, Ph.D. thesis, Florida Institute of Technology, 1997.

[4] Melissa Barkat, "Identification of Arabic dialects and experimental determination of distinctive cues," in *Proc. ICPhs-99*, August 1999, vol. 2, pp. 901–904.

[5] David R. Miller and James Trischitta, "Statistical dialect calssification based on mean phonetic features," in *Fourth International Conference on Spoken language Processing, Proceedings ICSLP 96*, 1996, vol. 4, pp. 2025–2027.

[6] Marwan Al-Zabibi, *An Acoustic-Phonetic Approach in Automatic Arabic Speech Recognition*, Ph.D. thesis, Loughborough University of Technology, 1990.

[7] M. H. Bakalla, *Arabic Culture Through its Language and Literature*, Kegan Paul, London, 2002.

[8] Alan S. Kaye and Judith Rosenhouse, *Arabic Dialects and Maltese in The Semitic Languages*, Robert Hetzron, Routledge, London, 1998.

[9] Salman H. Al-Ani, *Arabic Phonology: An Acoustical and Physiological Investigation*, Mouton, The Hague, Paris, 1970.

[10] Bruce Ingham, *Najdi Arabic : central Arabian*, Amsterdam, Philadelphia : John Benjamins Pub. Co., 1994.

[11] Theodore Prochazka Jr., *Saudi Arabian dialects*, London, New York : Kegan Paul International : Distributed by Routledge, Chapman and Hall, 1988.

[12] Imed Zitouni et. al., "Orientel: Speech-based interactive communication apllications for the mediterranean and the middel east," in *International Conference on Spoken Language Processing, ICSLP*, September 2002.

[13] Rainer Siemund et. al., "Orientel Arabic speech resources for the IT market," in *Third International Conference on Language Resources and Evaluation, LREC*, May 2002.

[14] Kees Versteegh, *The Arabic Language*, New York : Columbia University Press, 1997.

[15] Melissa Barkat, John Ohala, and Francois Pellegrino, "Prosody as a distinctive feature for the discrimination of Arabic dialects," in *Proc. Eurospeech 2000*, 1999, vol. Budapest, pp. 395–398.

[16] Marc A. Zissman, Gleason T.P., and B.L Rekart D.M.and Losiewicz, "Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech," in *Proc. IEEE ICASSP 96*, May 1996, vol. 2, pp. 777–780.

[17] Shuichi Itahashi, Du Liang, and Tanaka Kimihito, "A method of classification among spoken languages and dialects," 1995, available at `http://citeseer.ist.psu.edu/182370.html`.

[18] Shuichi Itahashi, Tanaka Kimihito, and Zhou Jiang Xiong, "Discrimination of spoken languages and dialects," 1994, available at `http://citeseer.ist.psu.edu/213183.html`.

[19] S. A. SantoshKumar and V. Ramasubramanian, "Automatic language identification using ergodic-HMM," in *Proc. IEEE ICASSP 2005*, March 2005, pp. 609–612.

[20] Marc A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proc. IEEE ICASSP 93*, April 1993, pp. 399–402.

[21] Joseph Picone, "Modeling techniques in speech recognition," in *Proceeding of the IEEE*, September 1993, vol. 81 No. 9, pp. 1215–1246.

[22] Lawrence Rabiner and Bing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall , New Jersey, 1993.

[23] Jan Cernocky, Pavel Matejka, Luka's Burget, and Petr Schwarz, "Automatic language identification system," in *Brno University of Defence*, 2006, pp. 1–6.

[24] Joseph P. Campbell JR., "Speaker recognition: A tutorial," in *Proceeding of the IEEE*, September 1997, vol. 85 No. 9, pp. 1437–1462.

[25] Lawrence R. Rabiner and Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall , New Jersey, 1978.

[26] David A. Nelson, Dianne J. Van Tasell, Anna C. Schroder, Sigfrid Soli, and Samuel Levine, "Electrode ranking of "place pitch" and speech recognition in electrical hearing," in *The Journal of the Acoustical Society of America*, 1995, vol. 98, pp. 1987–1999.

[27] Douglas A. Reynolds and Richard C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Seepch and Audio Proc.*, vol. 3, No. 1, pp. 72–83, January 1995.

[28] Steve Young et.al., *The HTK Book version 3.2.1*, Cambridge University Engineering Department, 2002.

[29] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” June 1976, vol. 55 No. 6, pp. 1304–1312.

[30] Steve Young, “Large vocabulary continuous speech recognition: A review,” in *Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding*, December 1995, vol. 77 No. 2, pp. 3–28.

[31] Douglas Alan Reynolds, *A Gaussian Mixture Modeling Approch to Text-Independent Speaker Identification*, Ph.D. thesis, Georgia Institute of Technology, 1992.

[32] Jeff A. Bilmes, “A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” Tech. Rep. ICSI-TR-97-021, University of Berkeley, 1997.

[33] Lawrence R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” in *Proceeding of the IEEE*, February 1989, vol. 77 No. 2, pp. 257–286.

[34] L.R. Rabiner and B.H. Juang, “An introduction to hidden markov models,” *IEEE ASSP Magazine*, vol. 13, No. 1, pp. 4–16, January, 1986.

[35] Joseph Picone, “Continuous speech recognition using hidden Markov models,” July 1990, vol. 7 No. 3, pp. 26–41.

[36] X.D. Huang, Y. Ariki, and M.A. Jack., *Hidden Markov models for speech recognition*, Edinburgh, Edinburgh University Press, 1990.

[37] Frederick Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts, 1999.

[38] “Linguistic data consortium,” available at `http://www.ldc.upenn.edu/`.

[39] F. Walls and et. al., “Probabilistic models for topic detection and tracking,” in *Proc. IEEE ICASSP 1999*, March 1999, pp. 521–524 vol.1.

[40] “Hidden markov models tool kit,” available at `http://htk.eng.cam.ac.uk/`.

[41] Igor Szke, “Speech units automatically generated by ergodic hidden Markov model,” in *Proceedings of 10th Conference and Competition STUDENT EEICT 2004*, 2004, p. 5.

[42] Pavel Matejka, Igor Szke, Petr Schwarz, and Jan Cernock, "Automatic language identification using phoneme and automatically derived unit strings," in *Proceedings of 7th International Conference Text,Speech and Dialoque 2004*, 2004, p. 8.

[43] George Casella and Roger L. Berger, *Statistical Inference*, Pacific Grove, Calif. : Brooks/Cole Pub. Co., 1990.

[44] Rupert G. Miller, "The jackknife-a review," in *Biometrika*, 1974, vol. 61, pp. 1–15.

[45] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, and J.R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *International Conference on Spoken Language Processing, ICSLP*, September 2002.

[46] M.A. Kohler and M. Kennedy, "Language identification using shifted delta cepstra," in *The 45th Midwest Symposium on Circuits and Systems, MWSCAS-2002*, August 2002, vol. 3, pp. III– 69–72.

[47] Fawzi Alorifi, "Arabic speech recognition: Review and implementation using htk," Preliminary Exam Proposal, University of Pittsburgh, Pittsburgh, PA, 2001.

[48] J. Deller, J. Proakis, and J. Hansen, *Discrete Time Processing of Speech Signals*, New York: Macmillan Publishing, 1993.

[49] Abdulhadi Al-Otaibi, *Arabic Speech Processing: Syllabic Segmentation and Speech Recognition*, Ph.D. thesis, The University of Aston, 1988.