Cairo University
Institute of Statistical Studies and Research
Department of Computer and Information Sciences

BUILDING A SPEECH RECOGNITION SYSTEM FOR SPOKEN ARABIC

by

Tarek Ahmed Fouad Ibrahim Sheisha

Under the Supervision of

Prof. Atef M. A-Moneim	**Prof. Khaled Fouad Shaalan**
Institute of Statistical Studies and Research Cairo University	Faculty of Computers and Information Cairo University

A Thesis Submitted to the
Department of Computer and Information Sciences
In Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
In
Computer Science

Cairo
August 2009

Approval Sheet

BUILDING A SPEECH RECOGNITION SYSTEM FOR SPOKEN ARABIC

By

Tarek Ahmed Fouad Ibrahim Sheisha

A Thesis Submitted to the Department of Computer and Information Sciences, Institute of Statistical Studies and Research, Cairo University in Partial fulfillment of the Requirements for the Degree of Master of Science in Computer Science.

Approved by The Examining Committee:

Name	Signature
Prof. Atef M. A-Moneim	..................Atef.
Dr. Hesham N. Elmahdy	......Hesh........................
Dr. Hesham A. Hefny	....Hesham...........

Cairo
August, 2009

Statement

I certify that this work has not been accepted in substance for any academic degree and is not being concurrently submitted in candidature for any other degree.

Student Name: Tarek Ahmed Fouad Ibrahim Sheisha

Signature: Tarek A. Fouad

Acknowledgements

I would like to express my sincere gratitude to Prof. Atef M. A-Moneim, who supervised and supported my work. This thesis was not made possible without his help and continuous support.

I am grateful to my advisor Prof. Khaled Shaalan for providing me guidance and support in my research.

I would also like to thank Prof. Lori Levin and Prof. Tanja Schultz (CMU) for their insightful discussions.

Finally, I would like to thank my family. They have greatly contributed to the inspiring, creative, and encouraging environment in which I have carried out my work toward this thesis.

ABSTARCT

The automatic recognition of spontaneously spoken speech is a very challenging task because it contains much more disfluencies than the read speech. The difficulty is even increased in case of recognizing colloquial speech. The colloquial speech contains many changes from the standard language such as changes in the syntax, the structure, the letters, and the diacritics. Moreover, the colloquial speech may contain different expressions and slang. Building of Arabic speech system is useful for adding user-friendly interfaces for Arabic software. Also, the embedding of the Arabic speech recognition system in the machine translation systems can allow human-to-human communication across languages boundaries.

In this thesis, a system was built for recognizing spontaneous Egyptian colloquial Arabic. A speech database was collected by recording 20 hours of spontaneously spoken Arabic dialogues. All spoken dialogues were segmented into separate utterances, the corresponding transcriptions were written. The diacritics (FatHa, Dammah, and Kasrah) are included in the transcription of each utterance. The speech database was used to train the acoustic models and the language model, which were used as knowledge sources for building the system. A baseline dictionary was built. A set of phones was chosen, which is suitable for the phonetics of Egyptian colloquial Arabic. A statistical language model, a word bigram Language model, was built using the canonical-form transcriptions of utterances of the speech database. A statistical acoustic model was built for each phone. The acoustic models were based on the Hidden Markov Models (HMMs). There are separate HMMs for the diacritics. All the acoustic models were trained using the speech database. Another system was built through training the models with non-diacritized transcriptions (i.e. the short vowels were removed).

The recognition experiments revealed that the system which was built with the fully diacritized transcriptions had a higher percentage of correctly recognized words. The system was modified by converting the monophone units into triphones units. Another modification was to increase the number of Gaussian mixtures associated with each HMM's state. The optimum values of the word insertion penalty P, and the grammar scale factor S were experimentally found. The percentage of correctly recognized words for the baseline system is 82.87 %.

A proposed data-driven method of modeling pronunciation variation is utilized to improve the baseline system. This method modifies the three levels of the speech recognition system at which modeling can take place; i.e. the dictionary, the acoustic model, and the language model.

A dynamic programming-based algorithm was developed to align the canonical-form transcription of each utterance with the string of phones constituting that utterance. For every utterance, a string of the constituting phones was obtained by a developed phone recognizer. As a result of the alignment process, a set of pronunciation rules was generated. The set of rules was applied on the whole set of canonical-form transcriptions to generate the surface-form transcriptions. The acoustic and language models were re-trained with the surface-form transcriptions. Then, by applying suitable rules on each entry in the dictionary, some pronunciation variants are generated. The dictionary was also modified by adding the pronunciation variants to the baseline dictionary.

The recognition experiments showed that the modeling of the pronunciation variations improved the percentage of correctly recognized words to 91.9 %.

A user interface was developed, through which the recognized sentences are displayed in diacritized Arabic scripts.

Contents

Acknowledgements …………………………………………………….. iii

Abstract ……………………………………………………...…. iv

Contents …………………………………………………..…….. vi

List of Figures ………………….………………………………….. xi

List of Tables ……………………..………………………………….. xiii

List of Algorithms ………………………………………………… xiv

Chapter 1 Background and Related Works

1.1 Introduction ………………………………………………. 1

1.2 Speech Recognition System Architecture ……………………. 2

1.3 The Source-Channel Model of Speech Recognition …...…….. 3

1.4 A Historical Background on Speech Recognition …………… 4

1.5 Thesis Goal ………………….………………………………. 5

1.6 Thesis Structure ……………………………………………... 6

Chapter 2 Automatic Speech Recognition

2.1 Introduction …………………………………………………... 7

2.2 Acoustic Front-End ………….………………………………. 8

2.2.1 Speech Signal ……………………………………….. 9

2.2.2 Acoustic Preprocessing …………………………….. 11

2.3 Acoustic Modeling …………………………………………… 16

2.3.1 The Role of Acoustic Modeling …...……………….. 17

2.3.2 Markov Chains …….……………………………….. 17

2.3.3 Hidden Markov Models ……………….…………….. 18

2.3.4 The Basic Problems of HMM …….……………….. 21

2.3.5 Evaluation of the HMM ………………………….. 22

2.3.6 The Forward Algorithm ……………..…………….. 23

2.3.7 Decoding HMMs ………………….……………….. 25

2.3.8 The Viterbi Algorithm ….................................. 25

2.3.9 Estimating of HMM Parameters ………………..….. 28

2.3.10 HMM Topology ………..……………………….. 31

2.3.11 Output Distributions …………………...……….. 32

2.3.12 Parameter Tying ………..……………………….. 34

2.3.13 The Need for a Language Model ………………….. 36

2.4 Language Modeling ………………………………………….. 37

2.4.1 Training the Language Model ……………...……….. 40

2.4.2 Perplexity of the Language model ………………….. 40

2.5 Applications of Speech Recognition Systems ……………….. 41

2.6 Summery …………………………………………………….. 42

Chapter 3 Search Strategies in Speech Recognition Systems

3.1 Introduction ……………….………………………………….. 43

3.2 The Basic Tasks of the Decoder …………….……………….. 45

3.3 Isolated-Word Recognition …………………..……………….. 45

a) Template-Matching Methods ……………..……………….. 45

b) Dynamic Time Warping DTW …………..……………….. 46

3.4 Continuous Word Recognition ………………………….……. 47

3.4.1 The One-Stage Dynamic Programming Algorithm … 48

3.5 Token-Passing Algorithm ……………………………………. 52

3.6 Decoding for the Large Vocabulary Continuous Speech Recognition ………………………………………………….. 53

3.7 Linear and Tree Lexicons ……………………………………. 54

3.8 Pruning Techniques ………………………………………….. 56

3.9 Cross-Word Expansion ………...…………………………….. 57

3.10 Single-Best versus N-best and Word Graphs ……………….. 58

3.11 Summery ………………………………...……………….. 58

Chapter 4 Arabic Speech

4.1 Introduction …………………………………….…………….. 59

4.2 Arabic Language …………………………………………….. 59

4.3 Egyptian Colloquial Arabic ………………………………….. 61

4.4 Speech Sounds …………………………………………………….. 61

4.5 Arabic Phonetics …………………………………………………. 66

4.5.1 Classifications of Arabic Phonemes and IPA …..….. 66

4.6 Arabic Alphabet and Writing System …………….……….. 68

4.7 Arabic Diacritics …………………………………………………. 70

4.8 Differences between MSA and the Egyptian Colloquial Arabic ……………..……………..……………………………. 72

4.9 A Survey on Arabic Speech recognition ………...…………… 74

4.10 Summery ………………….…..……………………………….. 77

Chapter 5 Baseline System Development and Experimental Results

5.1 Introduction …………………….…………………………….. 78

5.2 The Hidden Markov Model Toolkit …………..…………….. 79

5.2.1 The HTK Software Architecture ………...………….. 79

5.2.2 Overview on the Different Tools …….……………….. 81

5.3 Data Preparation Stage …………….……………………….. 85

5.3.1 Data Collection ……………………...…………….. 85

5.3.2 Recording Conditions …………….………………….. 85

5.3.3 The Dialogues ……………………….………………….. 86

5.3.4 Transcription of the Dialogues …………...…….. 86

5.3.4.a Transcriptions Standards ……………..………….. 87

5.3.4.b Transcription and Segmentation Software …..….. 87

5.3.4.c Creating the Transcription Files ……….…..…….. 88

5.3.5 Preparation of Pronunciation Dictionary ………….. 90

5.3.6 Extraction and Preprocessing of the Feature Vectors 91

5.3.7 Selection of the Topology and Prototypes of HMM 92

5.4 Training Stage ………………...…………………………….. 94

5.4.1 Training of the Acoustic Model …………...……….. 94

5.4.2 Training of the Language Model …………..……….. 96

5.5 Recognition Experiment 1……………….…………………….. 97

5.6 System Improvement by Increasing the Number of Components of Mixtures ……………………..…………….. 99

5.7 Recognition Experiment 2 …………………………………….. 100

5.8 Building a Context-Dependent System ………….………….. 102

5.8.1 Context-Dependent Models ……………………….. 102

5.8.2 State Tying ……………..………………………….. 103

5.8.3 Clustering Techniques …………………………….. 104

5.8.3.a Data-Driven Clustering …….…………………….. 104

5.8.3.b Decision Tree-Based Clustering ………………….. 105

5.9 Recognition Experiment 3 …………………………………….. 106

5.10 Recognition Experiment 4 ………..…………………….. 110

5.11 Recognition Experiment 5 ………..…………………….. 111

5.12 Summery ……………………...……………………….. 112

Chapter 6 System Improvement by Modeling the Pronunciation Variation

6.1 Introduction ………………………..…………………….. 113

6.2 Sources of the Pronunciation Variation …………...……….. 113

6.3 Methods of Modeling the Pronunciation Variation ………... 114

6.4 Pronunciation Variation in Conversational Egyptian Arabic.. 115

6.5 Canonical Transcription versus Surface-Form Transcription.. 116

6.6 Data-Driven Modeling of Pronunciation Variation ……..….. 116

6.7 Automatic Extraction of the Pronunciation Rules ………….. 117

6.7.1 Building of a Phone Recognizer …………...……….. 118

6.7.2 Developing an Algorithm for Automatic Rules Extraction ……………..……………………….. 119

6.8 Generation of Pronunciation Variants and Re-training of the System ……………………………….…………….. 121

6.9 Recognition Experiment 6 …………………….………….. 123

6.10 Recognition Experiment 7 …………………………………….. 125

6.11 Recognition Experiment 8 …………………………………….. 125

6.12 Recognition Experiment 9 …………………………………….. 127

6.13 Developing a User Interface ……………..…………………….. 128

6.14 Summery and Discussion ………………………..…….. 129

Chapter 7 Conclusion and Future Work

7.1 Conclusions ……………………….……………………….. 131

7.2 Future Work ……..…………..…………………………….. 133

Appendix A Characters Romanization Conventions for Egyptian Arabic ……………………..………………………………….. 134

Appendix B An Example from the Dialogues………...……………….. 135

Appendix C The Top 50 Pronunciation Rules ……………..………….. 136

Appendix D List of Abbreviations …….………………….……..…….. 137

References ………………………………………………..….. 138

List of Figures

Figure 1.1 Main Components of the Automatic Speech Recognition System 2

Figure 1.2 Source-Channel Model for Speech Recognition System 3

Figure 2.1 The waveform, the wideband spectrogram, and narrowband spectrogram of the phrase "بالتأكيد" (for sure).. 10

Figure 2.2 Block diagram of MFCC front-end analysis 13

Figure 2.3 Extracting the feature vectors from the speech signal 14

Figure 2.4 Mel-scale filterbank 14

Figure 2.5 The forward trellis computation for the HMM 24

Figure 2.6 The Viterbi trellis computation for the HMM 27

Figure 2.7 Operations required for the computation of $\gamma_t(i,j)$ 30

Figure 2.8 A typical Hidden Markov Model used to model a phoneme 32

Figure 3.1 The Alignment Path of the Dynamic Time Warping Algorithm.. 47

Figure 3.2 Search Space for the Continuous Word Recognition 49

Figure 3.3 Withen-Template Transition Rules 50

Figure 3.4 Between-Templates Transition Rules 51

Figure 3.5 An example of simple linear lexical search space 54

Figure 3.6 Example of a lexical tree 55

Figure 3.7 Cross-word triphone expansion network 58

Figure 4.1 Representation of the human speech production apparatus 62

Figure 5.1 The stages of speech recognition system development 79

Figure 5.2 HTK software architecture...................................... 80

Figure 5.3 A screen-snapshot of TransEdit software 88

Figure 5.4 An example of two label files associated with the same speech file 89

Figure 5.5 An excerption from a dictionary ………………………… 90
Figure 5.6 The configuration parameters for the feature extraction process …………………………………………………… 91
Figure 5.7 The HMM prototype …………………………………….. 92
Figure 5.8 A script that defines the HMM prototype ……………….. 93
Figure 5.9 The HMM of the silence model …………………………. 94
Figure 5.10 The HMM of the short pause model …………………..…… 94
Figure 5.11 A Gaussian mixture represented as a set of single Gaussians ………………………………………………… 100
Figure 5.12 Generation of a multiple mixture monophone model …… 101
Figure 5.13 Decision tree-based state tying ………………………….. 105
Figure 5.14 Experimental steps to generate triphone models ………… 107
Figure 5.15 Experimental steps to generate tied-state triphones …..… 110
Figure 6.1 An example of the input and output of the rules extraction algorithm ……………………………………...…… 121
Figure 6.2 The variation of system performance with the number of mixture components ……………………..……………… 126
Figure 6.3 The variation of the recognition time divided by the tested time duration with the number of mixture components….. 127
Figure 6.4 A graphical user interface of the recognition system …… 128
Figure 6.5 Comparison of the performance of different systems …… 129
Figure 6.6 Comparison of the word accuracies of different systems **..** 130

List of Tables

Table 4.1 Some Differences of Pronunciation of Phones………….. 60

Table 4.2 The International Phonetic Alphabet ……………………. 67

Table 4.3 Phonemes of Modern Standard Arabic (MSA)………….. 68

Table 4.4 Arabic Diacritics …………………………….……….. 69

Table 4.5 Arabic Alphabet ……………………………………… 71

Table 4.6 Examples of phonetic changes …………………………. 72

Table 4.7 Examples of changes in the diacritics …………………... 73

Table 4.8 Examples of changes in the vocabulary ..………………. 73

Table 4.9 Examples of changes in word order …………………….. 73

Table 5.1 The recording conditions ………………………………... 86

Table 5.2 Recognition Results for the Recognition Experiment 1 … 99

Table 5.3 Recognition Results for the Monophone Multiple-Mixture System ………………………………………. 102

Table 5.4 Some example questions used in building the phonetic tree ……………………………………………………. 108

Table 5.5 Recognition Results for the Tied-State Triphone system... 109

Table 5.6 Recognition Results for Multiple Mixture Triphone Systems 111

Table 5.7 Changing the word insertion penalty P, and the grammar Scale factor ………………………………….. 112

Table 6.1 Some Examples of the pronunciation Variation ………… 116

Table 6.2 Example of the output of the phone recognizer ………… 118

Table 6.3 Some examples of the extracted rules …………………... 119

Table 6.4 Examples of Multiwords ………………………………... 123

Table 6.5 Results of the Recognition Experiment 6 ……………….. 124

Table 6.6 Results of the Recognition Experiment 7 ……………….. 125

Table 6.7 Recognition Results Recognition Experiment 8 ………... 126

Table 6.8 Results of the Recognition Experiment 9 ……………….. 128

List of Algorithms

Algorithm 2.1 The Forward Algorithm ……………………………… 24

Algorithm 2.2 The Viterbi Algorithm ……………………….……… 26

Algorithm 2.3 The Forward-Backward Algorithm ………………...... 29

Algorithm 3.1 The One-Stage Dynamic Programming Algorithm …. 52

Algorithm 6.1 A Dynamic Programming-Based Algorithm to extract pronunciation rules ………………..……...….. 120