Person-Dependent and Person-Independent Arabic Speech Recognition System

Noor Al-Maadeed and Somaya Al-Maadeed

1 Introduction

Conversational interfaces have greatly improved over the last 20 years, leading to a new era of human–computer interaction. Voice recognition has recently begun to play an important role in information retrieval methods. However, while voice recognition technology for English speakers has been under investigation for more than 30 years, only a few studies have been performed on Arabic sounds.

Sounds can typically be recognized through the unique characteristics of each word. However, the pronunciation of a given word varies in a complex fashion among individuals, similar to music. Every person's voice is unique, and the same words can have different meanings depending on tone or perspective. Several approaches have been taken to solve this problem.

There are two popular frameworks for speech recognition: hidden Markov models (HMMs) [1], and artificial neural networks (ANNs) [2]. HMMs are simpler, faster, and generally require less training data than ANNs for most recognition tasks. In addition, HMMs can characterize the speech signal in a mathematically tractable fashion [3]. A hybrid HMM–ANN has recently been proposed to improve the performance by combining the two modeling strategies [3, 4].

Template matching is often used in audio classification problems. However, straightforward template matching often fails [5]. In [6, 7], continuous Gaussian mixture density models (CDHMMs) are shown to have error rates at least comparable to the rates of the best template recognizers and significantly lower than the

N. Al-Maadeed · S. Al-Maadeed (✉)
Department of Computer Science and Engineering, Qatar University, Doha, Qatar
e-mail: n.alali@qu.edu.qa; s_alali@qu.edu.qa

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_16

error rates of discrete symbol HMMs. This work focuses on the CDHMM approach because of its proven reliability and clear formulation.

In this chapter, the Arabic spoken word recognition problem is modeled using an HMM. The HMM states are identified with the sounds of the letters of the alphabet. Once the model is established, the Viterbi algorithm is used to recognize the sequence of letters composing the word.

2 Unique Characteristics of Arabic Words

The attributes, functionalities, and limitations of Arabic (or other non-Latin) speech recognition have been investigated previously. However, little attention has been devoted to the specific characteristics of Arabic word pronunciation. "Arabic is a Semitic language with approximately 221 million speakers in the Arab world and some African and Asian countries such as Chad, Cyprus, Iran, Israel, Kenya, Mali, Niger, Tajikistan, Tanzania, etc. ..." [8]. In addition, there are over 30 different varieties of colloquial Arabic.

English words are dissimilar to Arabic words in several ways, the most obvious and significant being the alphabet and symbols. There is no IPA (International Phonetic Alphabet) for the pronunciation of Arabic letters such as ع, ص، ض، ظ، ط and خ. Alghamdi Mansour (Mansour) has introduced a solution to this problem. Another distinct characteristic of Arabic words is their rhythm. The vowel and consonant lengths can affect the meanings of Arabic words, which is not the case in English. Even the most accurate voice recognition application does not reach 100% accuracy in Arabic. In this chapter, we study the ten Arabic numbersسبعة ثمانية, ستة, خمسة, ثلاثة اربعة, اثنين, تسعة واحد and وعشرة which include the unique Arabic letters ح, ع, خ ث. For multiple users, we studied the most popular Arabic words ان, من, ما في, مع, عن, الذي, التي, الى, على, most of which contain the letter ع. We then examined the most frequently used English words ("the," "of," "and," "in," "to," "was," "it," "is," "for," and "that"), omitting one-letter words such as "I."

3 System Overview

Our speech recognition system operates in three stages: feature extraction, generating the HMM model, and testing the model. The underlying model was a word-based HMM, as illustrated in Fig. 1. The following sections describe the training and testing of the model, including a discussion of the sound features used in the system. The HMM classifier, which classifies the features captured from the word image, is discussed in Sect. 4. The experimental results are presented in Sect. 5.

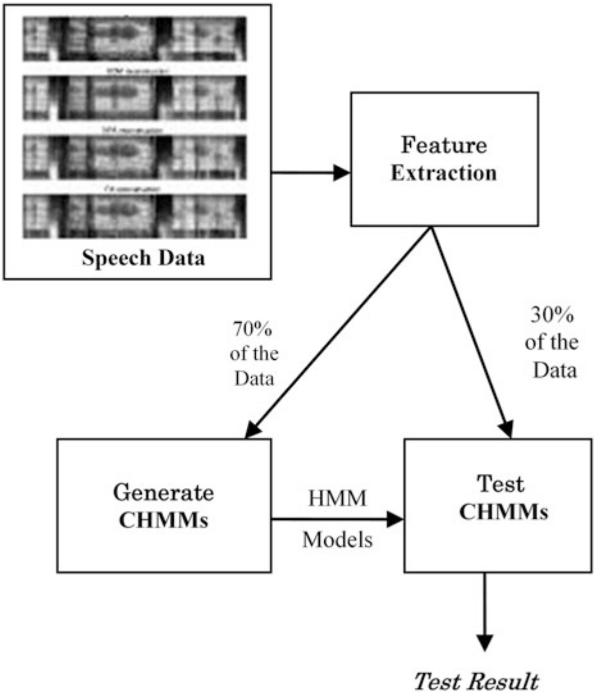

Fig. 1 The Arabic speech recognition procedure using the CDHMM

4 Data Content Preparation

In this stage, all of the data required by the system is collected, arranged in a suitable format for later use and stored. Three ten-word vocabularies are considered. The first vocabulary consists of the ten Arabic digits (TAD) between one and ten, as shown in Table 1. A word recognition test is performed by collecting the word vocabulary from ten different speakers. The Arabic numbers are used because many existing databases include a spoken digit. We recorded 100 wave sound files for ten different people, each of whom were asked to repeat each Arabic digit ten times.

The second vocabulary consists of the ten most popular Arabic words (MPAW), a list selected based on a prior study in which the number of occurrences of 30,000 words was counted on various Arabic websites. The top ten most common words (على, ان, من, في م, مع, عن, الذي, التي, الى ,) were selected. The voice of a single speaker was recorded ten times for each of these ten words.

The third ten-word vocabulary consists of the most used English words (MUEW) based on the whole corpus (spoken and written English): the, of, than, in, to, was, it, is, to, and for. The same speaker voice was recorded ten times for each of the ten most popular words in Arabic and English.

Table 1 The ten Arabic digits (TAD) between one and ten, with their meanings in English and transliterations [9]

	Arabic word	Meaning in English	Pronunciation
1	واحد	One	*Waahed*
2	اثنين	Two	*Athneien or athnan*
3	ثلاثة	Three	*Thalatha*
4	اربعة	Four	*Arbaa*
5	خمسة	Five	*Kamsaa*
6	ستة	Six	*Sitaah*
7	سبعة	Seven	*Saba'a*
8	ثمانية	Eight	*Thmaniea*
9	تسعة	Nine	*Tisa'a*
10	عشره	Ten	*A'ashra*

5 Feature Extraction

The feature extraction procedure consists of converting the voice sample into a series of appropriate vectors that describe the "features" of the signal. Feature extraction is performed regularly via short-term spectral analysis, which divides the speech signal into short frames of fixed length that can be assumed to be independent over a short time interval. Each frame usually overlaps its neighbor. To avoid "spectral artifacts" caused by discontinuities at the frame endpoints, the data in each frame is convolved with a smooth window function. The Hamming window is the most widely used smoothing function in speech recognition.

In most speech recognition systems, the acoustic features of choice are the Mel Frequency Cepstral Coefficients (MFCCs) [10]. The MFCCs are robust, contain a wealth of information regarding the vocal tract configuration regardless of the source of the excitation, and can be used to represent all classes of speech sounds. Other features, such as the Perceptual Linear Predictive (PLP) [11] coefficients, may also be used. In [12], an interesting set of acoustic parameters was presented, and their association with specific phonetic features was thoroughly investigated. However, the implementations presented in [12] are generally based on MFCCs.

6 The HMM Classifier

A Markov model can be described as a system with a set of states and a set of transitions between states. Each transition has an associated probability, and the system proceeds from state to state based on the current state and the probability of the transition to a new state. The observability of the states determines whether a Markov model is hidden. In standard Markov models, the states are directly observable. HMMs have states that are not directly observable; instead, there is a set of possible observations for each state and a set of allowed state transitions, and the observations of any one state depend on the probabilities of the possible observations.

When HMMs are used in pattern recognition, a set of observations is provided as an input signal. The HMM classifier then attempts to decide which of a set of possible HMMs is most likely to generate the given set of observations. The classification system contains a number of HMMs, each corresponding to a category, and the class corresponding to the HMM that best reproduces the incoming signal is the category assigned to that signal. HMMs work well with sounds that vary in duration because the durational change can occur in a single state or across many states. An HMM can begin by following the signal state for state, jump back and forth between a few middle states for the duration of the sustained portion of the sound, and then follow the decay portion down state for state. This procedure models the sound more accurately than a template that must be stretched to fit the length of the input signal.

An HMM is a stochastic process with an underlying Markov process that can only be observed through another set of stochastic processes produced by the Markov process (the observations are probabilistic functions of the states). Let us assume that a sequence of observations, $O = (o1, \ldots, oT)$, is produced by the state sequence $Q = (q1, \ldots, qT)$, where each observation ot is from the set of M observation symbols $V = \{vk;\ 1 \leq k \leq M\}$, and each state qt is from the set of N states $S = \{si;\ 1 \leq i \leq N\}$. An HMM can be characterized by $\pi = \{\pi i\}$, where $\pi i = P(q1 = si)$ is the initial state probability; $\pi = \{a_{ij}\}$, $aij = P(qt + 1 = sj \mid qt = si)$ is the state transition probability; and $\pi = \{bj\,(k)\}$, where $bj\,(k) = P(ot = vt \mid qt = j)$ is the symbol probability. The following probability constraints (1) must be satisfied:

$$\sum_{i=1}^{N} \pi = 1; \quad \sum_{j=1}^{N} a_{ij} = 1 \forall i; \quad \sum_{k=1}^{M} b_k(b) = 1 \forall j. \tag{1}$$

The HMM is specified using the compact notation $\lambda = \{\Pi, A, B\}$. We have only 55 classes (rather than 120 classes corresponding to the set of Arabic characters) because the recognition system is implemented on a limited set of words (the Arabic numbers). The 55 letters or sub-letters of the alphabet are defined to be the states of our HMM, where the initial πi and the transition probabilities a_{ij} are computed as follows:

$$a_{ij} = \frac{\text{number of trans.from } l\,(q_i)\ \ to\ l\left(q_j\right)}{\text{number of transitions from } l\,(q_i)} \tag{2}$$

where the function l maps a state to its representative member of the alphabet. Our system has two phases, training and testing, as illustrated in Fig. 1.

A. **Training problem**

Given the training sequence $O = o_1, \ldots, o_T$, the model parameters, $\lambda = \{\pi, A, B\}$, were adjusted to maximize $P(O/\lambda)$. The Baum–Welch Algorithm was used to determine the optimization criterion for finding the Maximum Likelihood (ML). In general, the Baum–Welch algorithm provides satisfactory performance in the training of HMMs [13].

B. **Recognition phase**

A Modified Viterbi Algorithm (MVA) was used to solve the recognition problem in the recognition phase [14].

7 Experimental Results

This section analyzes the output of our system to determine how well it recognizes Arabic speech in multi-user and single-user environments. The performance in an Arabic environment is also compared to the performance in an English single-user environment. An evaluation of the continuous hidden Markov model implementation phase was performed to determine its score on Arabic words compared to English words. This section describes the results of these tests and the problems encountered. As discussed in Sect. 3, we study the recognition of the ten Arabic digits (TAD) for HMMs with different combinations of parameters in a multi-user environment and then for different users. The approach taken in this project is as follows: The system performance was evaluated based on two criteria: the recognition rate and how well each word is recognized compared to other possible incorrect solutions.

The same procedure was performed on the most popular Arabic words (MPAW) and the most used English words (MUEW). The recognition rate was calculated for each vocabulary, and the discrepancies among different possible solutions were analyzed. The results for the Arabic word list were then compared with the results for the English words.

The results can be divided into two main parts: the results for a multi-user environment and the results for individual users. The list of ten Arabic digits (TAD) is used for the first (multi-user) part of the analysis. Two aspects of the model were analyzed: first, the effect of the choice of model parameters (the number of states, number of mixtures, and number of iterations) and second, the word characteristics of the different models. The words were tested on different individuals and drawn if there was a character for the Arabic word.

The second part of the analysis involves individual user voice recognition for Arabic and English words. The system is tested on MPAW (the most popular Arabic words) and MUEW (the most used English words). Two quantities were calculated from the output conflict matrix described in Chap. 4: the recognition rate and the variation among the optimal solutions.

To determine the role of the individual character or other word in the evolution process and the system implementation, a wider-scale application of the system will be discussed. Suggestions for improving the evaluation and program are discussed at the end of this chapter along with the limitations of the project.

The recognition rate for the three databases is calculated using a combination of three variables: the number of states, the number of mixtures, and the number of iterations. This section discusses the TAD (ten Arabic digit database) performance in a multi-user environment. The details of the measurements and performance evaluations for multi-user TAD and for the individual-user MPAW and MUEW environments are discussed in the following subsections.

7.1 TAD Results for Multiple Users

The proposed system first loads files for each word from word "one" to word "ten." Ten samples are used for each speaker (for a total of 100 files) in the feature extraction to generate the HMMs, training model and test model. The overall recognition rate is shown in Table 2.

The CDHMM training model employed 70% of the voices, and the testing model used 30% of the voices. By analyzing the outcome of the test model, it was found that different results are obtained for different values of the variables (the number of states or number of iterations). The optimal number of states is seven or three, and the number of states yielding the worst performance is four. The number of iterations is optimal when the number of mixtures is equal to two or four and decreases as the more mixtures are added. This trend occurs because the training data for each HMM model consists of seven voices. Any number higher than seven is therefore inadmissible. The number of iterations decreases and then stabilizes to a constant value of one. All of the users are female, but two regional accents were detected. The speakers also vary in their degree of consistency. Some individuals are consistent in their speech; each time these individuals are asked to pronounce a word, they produce similar acoustic (acoustic) signals with only minor variations. Consistent speakers are referred to as "sheep," while inconsistent speakers are referred to as "goats" [14]. Speaker number 10 in the TAD sample was clearly a "goat" as the measurement for this speaker was always the closest to zero (had the smallest absolute value). The variance percentage in the solution for multiple users is −1.5.

The effects of individual variation were also studied by analyzing the HMM parameters for the multi-user Arabic dataset. Various sets of volunteers were examined and trained to determine how the differences among individual speakers might affect the results. In this analysis, the datasets used for the training and the speaker used in the model evaluation were varied. TAD1 was used as the training model for volunteers/speakers 1–7 and tested on the remaining volunteers

Table 2 Best recognition rates for a single user and for multiple users

Database	Recognition rate (%)	Variance measurement
TAD	84	−1.5
MPAW	83	−0.34
MUEW	76.6	−0.28

Table 3 Recognition rate of different users

Training data (model No)	Test no.	User providing test data	Recognition rate (%)	Required solutions variance
TAD1	1	User no. 8	50	−1.41
TAD1	2	User no. 9	70	−1.15
TAD1	3	User no. 10	70	–0.34
TAD2	4	User no. 1	80	−2.14
TAD2	5	User no. 2	90	−1.6
TAD2	6	User no. 3	90	−1.9
TAD3	7	User no. 3	70	−1.87
TAD3	8	User no. 6	90	−1.87
TAD3	9	User no. 9	80	−1.19
Average	10			−1.5

(8–10), while TAD2 was the training model for volunteers 4–10 and was tested on volunteers 1–3. TAD3 was also tested for every third volunteer in the row (volunteers 3, 6, 9) and trained on the remaining voices.

Table 3 demonstrates that different speakers produce different recognition rates. However, the recognition rates may also be affected by factors that have not been considered. For example, speakers 9 and 10 produce similar recognition rates (70%), but the other possible solution may approximate the required solution. In other words, some speakers may speak more clearly than others yet produce the same recognition rate. A measure of the deviation of the alternative solutions from the required one is therefore needed. A larger absolute value of the measure indicates a better result. Therefore, speaker 9 with TAD1 has a much clearer voice than speaker 10 with the same training data (TAD1), as shown by the measurement of 1.15 for speaker 9 compared to 0.34 for speaker 10.

Table 3 demonstrates that different speakers have different recognition rates. The highest recognition rates (90%) are obtained for speakers 2 and 3 with TAD2 and speaker 6 with TAD3. The lowest recognition rate (50%) was obtained for speaker 8. It has also been observed that a speaker may produce different recognition rates depending on the training data used. For example, the recognition rates for the third speaker vary from 90% (with the TAD2 model) to 70% (with TAD3); this means that the recognition rate depends not only on the tested user but also on the training voices. In addition, note that the pronunciation of the words is marginally similar (−1.87 and −1.9) and slightly above the average value, possibly because the speaker still speaks with the same clarity.

The same observation applies to speaker 9, who delivers a recognition rate of 70% and 80% using TAD1 and TAD3, respectively. In addition, the variance in the result is similar for both cases (−1.15 and −1.19), demonstrating that the tested voice outcome was not influenced by the training voice.

Table 3 also shows that speaker 8 produces a recognition rate of only 50%. Nevertheless, speaker 10 (with a recognition rate of 70%) may have inferior voice recognition compared to speaker 8 as the measurement in the case of a distinguishable solution is −0.34 compared to −1.4 for speaker 8 (see Table 4).

Table 4 Recognition rate for entire words (Number of states = 3, Number of mixtures = 3)

Data set	Average recognition rate (%)	Optimal solutions variance
TAD1	63	−1.11916
TAD2	76	−2.3454
TAD3	80	−2.45801

Table 5 Characteristics of each word in the TAD database

Word	Recognition rate (%)	Required solution variance × −20	Required solution variance
واحد	77.78	54.6	−2.73
اثنين	77.78	31	−1.55
ثلاثة	55.56	31.8	−1.59
أربعة	88.89	41.8	−2.09
خمسة	55.56	42.8	−2.14
ستة	55.56	50.2	−2.51
سبعة	88.89	44.6	−2.23
ثمانية	55.56	20.8	−1.04
تسعة	77.78	57.6	−2.88
عشرة	100	70.6	−3.53
All	73.33	30.2	−1.51

The third group (TAD3) delivers the highest recognition rate, while the recognition using the first group (TAD1) is lowest. This result occurs because speaker 9, with a poor recognition rate, was included in TAD1. The other groups achieve a high word recognition accuracy. The probability also increases as the voice recognition improves for the other groups. The words three and eight have low overall recognition rates, again because TAD1 includes an unclear speaker, especially for those two words. Both words include the letter “thaa ث.”

Table 5 summarizes the recognition results for every word in the TAD database. The word ten عشرة is always recognized, while the words three ثلاثة, five خمسة, six ستة, and eight ثمانية have a much lower recognition rate (55%). The word eight ثمانية is the least distinguishable (the solution variance is −1.08), followed by the words three, five, and six. The words four “اربعة “and seven “سبعة “have high recognition rates. All of the words with recognition rates of 88% or higher include the letter “ع” (“A’aa” in English) and have solution variances of −2 or larger.

7.2 *Analysis of the Results for Single Users on MPAW and MUEW*

Table 6 shows that words number 3, 4, 5, 8, 9, and 10 (عن, الى, على, ما ان, مع) are recognized without error. In the TAD list, only the words for 4, 5, 8, and 9 contain the letter A’a. The words for the numbers 3 and 8 include the letter (I’ian ع).

Table 6 Words ordered by variance

Word number	Word in Arabic	Optimal solution variance	Recognition rate (%)
4	على	−0.45	100
9	مع	−0.43	100
8	عن	−0.41	100
5	الى	−0.4	100
1	في	−0.39	33
6	التي	−0.36	66
7	الذي	−0.32	66
10	ما	−0.24	100
3	ان	−0.23	100
2	من	−0.14	66

Table 7 Analysis of the recognition rate for the ten most used words in written and spoken English (MUEW)

Word no	English word	Recognition rate (%)	Submission rate (%)	Most often conflicted with	Optimal solution variance
1	The	66	33	Word no. 10	−0.34
2	Of	66	33	Word no. 10	−0.29
3	And	66	66	Word no. 7, 8	−0.33
4	In	66	33 (small error rate)	Word no. 7	−0.30
5	To	66	33	Word no. 7	−0.26
6	Was	100	0		−0.30
7	It	100	0		−0.32
8	Is	100	0		−0.43
9	For	33	66	Word no. 6	−0.07
10	That	100	0		−0.18

Word number 6 التي is pronounced "alatee," and word number 7 الذي is pronounced "alathee." Because of their similar pronunciations, these words are often confused with each other. Word number 2 من is the most often confused; its solution variance measure is 0.14 as it can easily be confused with word numbers 1, 5, 6, and 7. The top three words include the letter A, ع, and the following four include the letter ى or ي (measurements of −0.4 to −0.32). The letters ن ا م (pronounced aa, m, and n) yield measurement values between −0.24 and − 0.14. Overall, the average variance in the solution was −0.36 for 90% of the words and − 0.34 for all of them (Table 7).

The words "was," "it," "is," and "that" were recognized without error, but the word "that" had a lower measurement variance (less than −0.08), while the remaining words had variances exceeding −0.25. The variance of the required 90% of the words is approximately −0.31.

In general, we can conclude that Arabic words have a higher recognition rate than English words due to the strong and unique sounds of Arabic characters such as “I’an ع.” Note that Arabic includes more sounds and characters than other languages.

8 Conclusion

In this chapter, a speaker-independent voice recognition system using continuous HMMs has been proposed, developed, and evaluated. After the data are collected, their features are extracted into ten different files, and ten models are trained and tested using maximum likelihood methods, as discussed in Sect. 6. The voice recognition system is trained to maximize the likelihood of a given speech pattern based on the testing environment. The experimental results on multiple users display a correct identification rate of 80% for the list of numbers. This recognition rate can be improved by post-processing, in which more training samples are added. The Arabic language includes a special letter or sound “ع” (“A’aa” in English) shared by other Semitic languages such as Hebrew. We have shown that using words containing the letter “ع” (“A’aa” in English) in a speech recognition system reduces the error rate. On the other hand, words containing soft sounds had lower recognition rates. In this work, relationships were identified between speech recognition and specific sounds in the Arabic language; these relationships will be investigated in further depth for speaker-independent speech recognition in future studies. We have established that the recognition rate depends not only on the tested user but also on the voices used for training.

Acknowledgments This publication was made possible by a grant #NPRP7-442-1-082 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine, 3*(1), 4–16.
2. Bengio, Y., De Mori, R., Flammia, G., & Kompe, R. (1992). Global optimization of a neural network-hidden Markov model hybrid. *IEEE Transactions on Neural Networks, 3*(2), 252–259.
3. Hifny, Y., & Renals, S. (2009). Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(2), 354–365.
4. Holmes, J., & Holmes, W. (2001). Neural networks for speech recognition. In J. Holmes & W. Holmes (Eds.), *Speech synthesis and recognition* (pp. 217–218): CRC Press.
5. Gerhard, D. (2007). *Audio Signal Classification: History and Current Techniques. University of Regina Technical Report TR-CS*

6. Juang, B., Rabiner, L., Levinson, S., & Sondhi, M. (1985). *Recent developments in the application of hidden Markov models to speaker-independent isolated word recognition.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'85.
7. Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall.
8. Juang, B. H., Hou, W., & Lee, C. H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing, 5*(3), 257–265.
9. Al-Ma'adeed, S., Elliman, D., & Higgins, C. A. (2000). A database for Arabic handwritten text recognition research. *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition, 2000, 8*, 130–135.
10. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing, 28*(4), 357–366.
11. Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America, 87*, 1738.
12. Hansen, A. V. (1997). Acoustic parameters optimised for recognition of phonetic features. *Proceedings of Eurospeech-97*, 397–400.
13. Bunke, H., & Wang, P. S. (1997). *Handbook of character recognition and document image analysis*: World Scientific Publishing Company Incorporated.
14. Doddington, G. R. (1998). Sheep, Goats, Lambs and Wolves - An Analysis of Individual Differences in Speaker Recognition Performance, from http://simson.net/ref/1998/Sheep_Goats_Lambs_and_Wolves.pdf
15. Mansour, A. Arabic Script Retrieved 6 June 2005, from http://www.omniglot.com/writing/arabic.htm