

# SPEAKER-INDEPENDENT NATURAL ARABIC SPEECH RECOGNITION SYSTEM

Moustafa Elshafei\*, Husni Al-Muhtaseb\*, and Mansour Al-Ghamdi\*\*

\*King Fahd University of Petroleum and Minerals, \*\*King Abdulaziz City of Science and Technology

[elshafei@kfupm.edu.sa](mailto:elshafei@kfupm.edu.sa), [muhtaseb@kfupm.edu.sa](mailto:muhtaseb@kfupm.edu.sa), [mghamdi@kacst.edu.sa](mailto:mghamdi@kacst.edu.sa)

**Keywords:** Arabic Speech Recognition, Natural Language, News Transcription, Sphinx training.

## Abstract

This paper reports the results of the first phase of a research work for building a high performance, speaker-independent natural Arabic speech recognition system. This work aims at developing an Arabic broadcast news transcription system and a base system for further research. Several concurrent recent advances in Arabic language processing were crucial for the success of this stage, e.g automatic generation of Arabic diacritical marks, and rule-based phoneme dictionary. The developed Arabic speech recognition system is based on the Carnegie Mellon university Sphinx tools. The Cambridge HTK tools were also utilized at various testing stages.

The engine uses 3-emitting states Hidden Markov Models (HMM) for triphone-based acoustic models. The state probability distribution uses continuous density of 8 Gaussian mixture distributions. The system was trained on 4.3 hours of the 5.4 hours of Arabic broadcast news corpus and tested on the remaining 1.1 hours. The phonetic dictionary contains 23,841 definitions corresponding to about 14232 words. The language model contains both bi-grams and tri-grams. The Word Error Rate (WER) came to 9.0%.

## 1 Introduction

The statistical approach for speech recognition [17,18] has virtually dominated Automatic Speech Recognition (ASR) research over the last few decades, leading to a number of successes [24,28]. The statistical approach is itself dominated by the powerful statistical technique called Hidden Markov Model (HMM). The HMM-based ASR technique has led to numerous applications requiring large vocabulary speaker-independent continuous speech recognition.

The HMM-based technique essentially consists of recognizing speech by estimating the likelihood of each phoneme at contiguous, small frames of the speech signal [24,25]. Words in the target vocabulary are modeled into a sequence of phonemes, and then a search procedure is used to find, amongst the words in the vocabulary list, the phoneme sequence that best matches the sequence of phonemes of the spoken words.

Each phoneme is modeled as a sequence of HMM states. In standard HMM-based systems, the likelihoods (also known as the emission probabilities) of a certain frame observation

being produced by a state is estimated using traditional Gaussian mixture models. The use of HMM with Gaussian mixtures has several notable advantages such as a rich mathematical framework, efficient learning and decoding algorithms, and an easy integration of multiple knowledge sources.

Two notable successes in the academic community in developing high performance large vocabulary speaker independent speech recognition systems are the HMM tools, known as the HTK tool kit, developed at Cambridge University [15]; and the Sphinx system developed at Carnegie Mellon University [16,20,22,27], over the last two decades.

The Sphinx tools can be used for developing wide spectrum of speech recognition tasks. For example, the Sphinx-II [16] uses the Semi-Continuous Hidden Markov Model (SCHMM) models to reduce the number of parameters and the computer resources required for decoding, but has limited accuracy and complicated training procedure. On the other hand Sphinx-III uses the Continuous Hidden Markov Model (CHMM) with higher performance, but requires substantial computer resources [22]. Sphinx-4, which was developed in Java, can be used for building platform independent speech recognition applications [19,27].

Development of an Arabic speech recognition system is a multi-discipline effort, which requires integration of Arabic phonetic [1,2], Arabic speech processing techniques [3,10], and Natural language [11,12]. Development of an Arabic speech recognition system has recently been addressed by a number of researchers. Recognition of Arabic continuous speech was addressed by Al Otaibi, [21]. He provided a speech dataset for Modern Standard Arabic (MSA). He studied different approaches for building the Arabic speech corpus, and proposed a new technique for labeling Arabic speech. He reported a recognition rate for speaker dependent ASR of 93.78% using his technique. The ASR was built using the HTK tool kit.

Bila et al. [6] addressed the problems of indexing of Arabic news broadcast, and discussed a number of research issues for Arabic speech recognition.

There are a number of other attempts to build Arabic ASR (AASR), but they considered either limited vocabulary, or speaker dependant system [4,8,9,23,26].

In this work, we utilized the state of the art speech recognition engines developed at Carnegie Mellon University CMU and Cambridge University to build a natural language, large vocabulary, speaker independent Automatic Arabic Speech Recognition (AASR) system.

The development of an Arabic speech recognition system requires in the first place the building of an Arabic Speech Corpus. The training of the speech recognition models requires also building an Arabic phonetic dictionary together with its management tools. The dictionary should contain all possible phonetic pronunciations of any word in the domain vocabulary. The next step is to build the acoustic model and the language model as it will be explained in the subsequent sections. In Section 2, we provide an overview of the various components of the Sphinx speech recognition engine. Then, in Section 3 we describe the training steps. Section 4 covers the Arabic speech corpus, and the phonetic dictionary. Finally, in Section 5, we present an evaluation of the developed system and the recognition results.

## 2 System Description

Fig.1, illustrates the various components of the Sphinx speech recognition. The following is a brief description of the main components

**The Front-End:** This sub-system provides the initial step in converting sound input into a form of data usable by the rest of the system (called features). The front-end used in our analysis is the Mel-Frequency Cepstral Coefficients (MFCC).

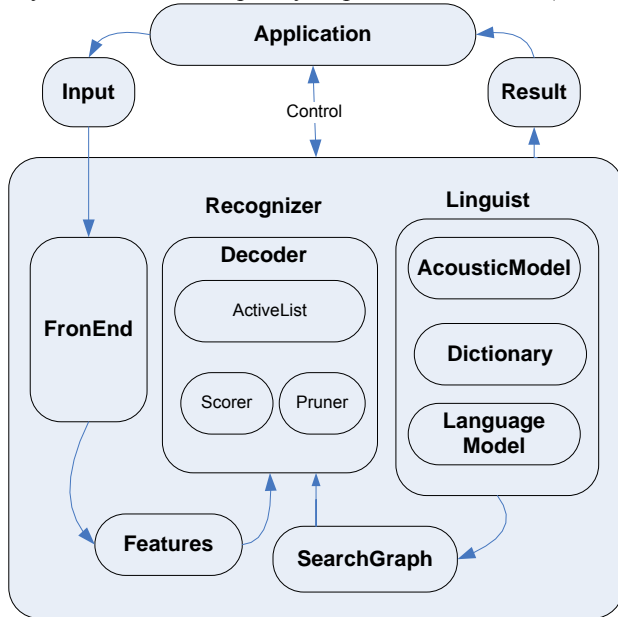


Fig. 1: Speech recognition system's architecture.

**The Linguist:** This sub-system contains the details that describe the recognized language itself. This sub-system is where most of the adjustments are made in order to support the Arabic Language recognition. It consists of three main modules:

*The Acoustic Model:* This module provides the Hidden Markov Models (HMMs) of the Arabic triphones to be used to recognize speech.

*The Language Model:* This module provides the grammar that is used by the system (Usually the grammar of a natural language or a subset of it).

*The Dictionary:* This module serves as an intermediary between the Acoustic Model and the Language Model. It contains the words available in the language and the pronunciation of each in terms of the phonemes available in the acoustic model.

**The Decoder:** This sub-system does the actual recognition job. When speech is entered into the system, The Front-End converts the incoming speech into features as described earlier. The Decoder takes these features, in addition to the search graph provided by the Linguist, and tries to recognize the speech. The decoder requires, in addition to the acoustic models, the phonetic dictionary and the language model.

### 2.1 Feature extraction:

The recorded speech is sampled at a rate of 16 ksp. The analysis window is 25.6 msec (about 410 samples), with consecutive frames overlap by 10 msec. Each window is pre-emphasized and is multiplied by a Hamming window [27]. The basic feature vector uses the Mel Frequency Cepstral Coefficients MFCC. The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The MFCCs are obtained by taking the Discrete Cosine Transform (DCT) of the log power spectrum from Mel spaced filter banks [17]. 13 Mel frequency cepstra are computed,  $x(0), x(1), \dots, x(12)$ , for each window of 25 ms, with adjacent windows overlapped by 15 ms.  $x(0)$  represents the log mel spectrum energy, and is used to derive other feature parameters. The system uses the rest 12 coefficients as a basic feature vector. The basic feature vector is usually normalized by subtracting the mean over the sentence utterance. The basic feature vector is highly localized. To account for the temporal properties, 3 other derived vectors are constructed from the basic MFCC coefficients: a 40-ms and 80-ms differenced MFCCs (24 parameters), a 12-coefficient second order differenced MFCCs, and 3-dimensional vector representing the normalized power (log energy), differenced power, and second-order differenced power

### 2.2 Acoustic model

The HMM model to represent the speech phoneme is shown in Fig. 2, where  $a_{ij}$  are the state transition probabilities, and  $b_i$  are the emission probabilities. The model, known as Bakis model, has a fixed topology consisting of 3 emitting states, an input state, and one output state. The state emission probabilities use a Gaussian mixture density model. HMM with Gaussian Mixture probabilities are called Continuous Hidden Markov Model (CHMM).

The probability of generating the observation  $x_t$  given the transition state  $j$ ,  $P(x_t | j)$  becomes

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

Where  $N_{j,k}$  is the k-th Gaussian distribution,  $w_{j,k}$  are the mixture weights, and  $\sum_k w_{j,k} = 1$ . CHMM is the most effective method today for building large-vocabulary speech recognition applications.

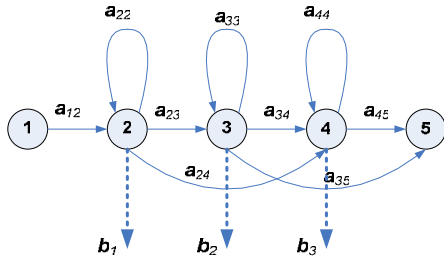


Fig.2: The HMM phoneme model.

The output probability distributions in HMM states are modeled with mixtures of 8 diagonal covariance Gaussian distributions.

### 2.3 Arabic Phoneme Set

Table 2 shows the listing of the phoneme set used in the training stage and the corresponding symbols. The table also shows illustrative examples of the vowel usage.

الرمز الصوتي	الحرف	الرمز الصوتي	الحرف
/AE/	◀ بَ	/KH/	خ
/AE:/	◀ بَاب	/D/	د
/AA/	◀ خَ	/DH/	ذ
/AH/	◀ قَد	/R/	ر
/UH/	◀ بُ	/Z/	ز
/UW/	◀ وُون	/S/	س
/UX/	◀ عُصن	/SS/	ص
/IH/	◀ بنت	/DD/	ض
/IY/	◀ فيل	/TT/	ط
/IX/	◀ صنف	/DH2/	ظ
/AW/	◀ لوم	/AI/	ع
/AY/	◀ ضيف	/GH/	غ
/UN/	◀ ننجي	/F/	ف
/AN/	◀ نم	/V/	ف◀ فيزا
/IN/	◀ مما	/Q/	ق
/E/	◀ ء	/K/	ك
/B/	◀ ب	/L/	ل
/T/	◀ ت	/M/	م
/TH/	◀ ث	/N/	ن
/JH/	◀ جيم فصحة	/H/	هـ
/G/	◀ جيم مصرية	/W/	و
/ZH/	◀ جيم معطشة	/Y/	ي
/HH/	◀ ح		

Table 1: The phoneme list used in the training.

## 3 Training Steps

Training the complete speech recognition engine consists of building two models, the language model and the acoustic model.

### 3.1 Acoustic Model Training

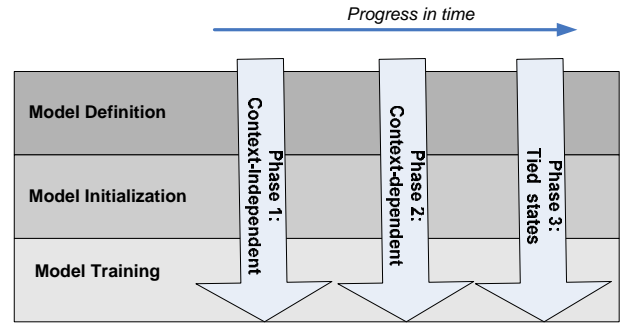


Fig. 3: Acoustic model-building steps

The training procedure consists of three phases as shown in Fig. 3. Each phase consists of three steps; model definition, model initialization, and model training. In the first phase, Context-Independent (CI) phoneme models are built. Baum-Welch re-estimation algorithm is used iteratively to estimate the transition probabilities of the CI HMM models [24,25]. In this phase the emission probability distribution of each state is taken to be a single normal distribution.

During the second phase, an HMM model is built for each triphone, that is a separate model for each left context and right context for each phoneme. During this context-dependant (CD) phase, triphones are added to the HMM set. In the model definition stage, all the possible triphones will be created, and then the triphones below a certain frequency are excluded. After defining the needed triphones, states are given serial numbers as well (continuing the same count). The initialization stage copies the parameters from the CI phase. Similar to the previous phase, the model training stage consists of iterations of the Baum-Welch algorithm (6 to 10 times) followed by a normalization process. The re-estimation is performed iteratively.

The number of tri-phones came to 10326. Table 2 below gives the number of tri-phones for each of the Arabic phonemes according to the current speech corpus, Version 1 (5.4 Hours).

Phone	Triphones	Phone	Triphones
AA	96	IX:	51
AA:	70	IY	372
AE	542	JH	181
AE:	389	K	225
AH	64	KH	130

AH:	40	L	560
AI	289	M	344
AW	77	N	454
AY	104	Q	238
B	324	R	460
D	356	S	302
DD	137	SH	144
DH	65	SS	156
DH2	41	T	393
E	479	TH	106
F	286	TT	161
GH	83	UH	487
H	258	UW	257
HH	195	UX	70
IH	657	W	187
IX	85	Y	218
IX:	51	Z	192

Table 2: Number of tri-phones for each phoneme in the AASR.

The performance of the model generated by the previous phase is improved by tying some states of the HMMs. These tied states are called Senons. In the third training phase, the number of distributions is reduced by combining similar state distributions. The process of creating these senons involves classification of phonemes according to some acoustic property. Decision trees are used to decide which of the HMM states of all the tri-phones (seen and unseen) are similar to each other, so that data from all these states are collected together and used to train one global state, which is called a senon. A senon is also called a tied-state and is obviously shared across the triphones which contributed to it. In the last phase, the senons probability distributions are re-estimated and presented by a Gaussian mixture model by iterative splitting of the Gaussian distributions. In this reported work, the emission probabilities of the senons are modeled with mixtures of 8 diagonal covariance Gaussian distributions.

### 3.2 Arabic Language Model

The steps for creating and testing the language model are shown in Fig. . The creation of a language model from a training text consists of the following steps:

- Compute the word unigram counts.
- Convert the word unigram counts into a task vocabulary with word frequencies.
- Generate a bi-grams and tri-grams from the training text, based on this vocabulary.
- Convert the n-grams into a binary format language model, and standard ARPA format.

The number of unigrams came to 14231, bigrams=32813, and trigrams=37771.

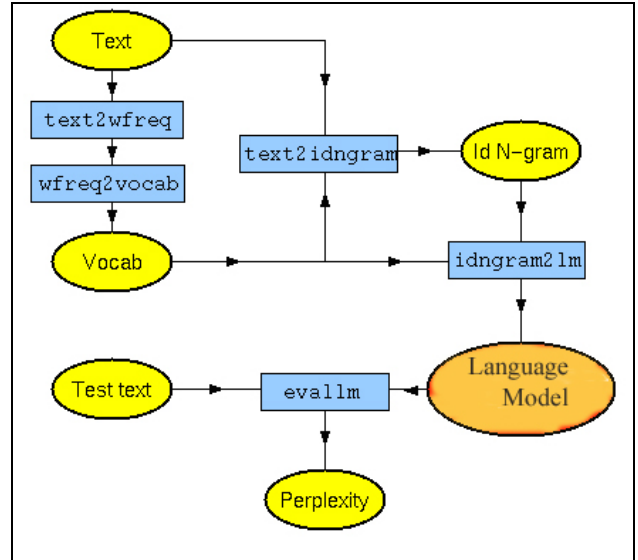


Fig. 4: Steps for creating and testing Language Model.

## 4 Arabic Broadcast News Corpus

The audio files were recorded from several TV news channels at a sampling rate of 16 ksp/s. A total of 249 news stories, summing up to 5.4 hours of speech, were recorded and split into 4572 files with an average file length of 4.5 seconds. The length of wave files range from 0.8 seconds to 15.6 seconds. Additionally, a 0.1 second silence period is added to the beginning and end of each file. Although care was taken to exclude recordings with background music or excessive noise, some of the files may still have background noise such as low level or fainting music, environmental noise when the reporter is in an open location such as a stadium or a stock market, and low level overlapping foreign speech, when the reporter is translating a foreign statement.

### 4.1 Transcription:

All the 4572 files were completely transcribed with fully diacritized text. The transcription is meant to reflect the way the speaker has uttered the words, even if it is grammatically wrong. It is a common practice in Modern Standard Arabic (MSA) and most Arabic dialect to drop the vowels at the end of words; this situation is represented in the transcription by either using a silence mark (Sukun) or dropping the vowel, which is considered the same as a Sukun in later training stages. The original transcription had some errors that required many corrections and revisions, including missing diacritical marks, miss-vocalized or miss-spelled words, and inconsistent in vocalizing “ ʔ ”, and Tanween following

Alef. The transcription file contains 39,217 words. The vocabulary list contains **14,231** words.

To speed up the diacritization, an algorithm was developed by the authors for automatic vocalization of the Arabic text. The detailed algorithm is given in [12,13]. The word sequence of undiacritized Arabic text is considered an observation sequence from an HMM, where the hidden states are the possible diacritized expressions of the words. The optimal sequence of diacritized words (or states) is then obtained efficiently using Viterbi Algorithm. However, the correct letter transcription came to about 90% since, the system was trained on different text subjects. Hand editing was then necessary to bring the transcription to the desired accuracy level.

#### 4.2 Arabic Phonetic Dictionary

Using the selected phoneme set, we developed a Java tool that automatically generates a dictionary for a given transcription. Automatic generation of Arabic pronunciation dictionary was recently addressed by Hiyassat in [14]. Hiyassat developed a tool kit for building a pronunciation dictionary for the Holy Quran, and for two other small corpuses, a 30 command corpus, and Arabic digits. On the other hand, the developed tool for our work is built for natural language MSA, and takes care of the following issues:

- 1- Choosing the correct phoneme combination based on the location of the letters and their neighbors using language pronunciation rules.
- 2- Providing multiple pronunciations for words that are pronounced in different ways according to:
  - a) The context in which the words are uttered, which might change the way of the pronunciation of the beginning and the end of the word. For example, Hamzat al-wasl (ا) at the beginning of the word and the Ta' al marbouta (ة) at the end of the word.
  - b) Words that have multiple readings due to dialect issues.
  - c) Common foreign names, such as "Lagrange", "Vector", etc., where the translation might not reflect the exact pronunciation.

We defined a set of rules based on regular expressions to define the phonemic definition of words. The tools scans the word letter by letter, and if the conditions of a rule for a specific letter are satisfied, then the replacement for that letter is added to a tree structure that represents all the possible pronunciations for that words. The number of pronunciations in the developed phonetic dictionary came to 23840 entries. A sample from the developed phoneme dictionary is listed below.

أَبَار E AE: B AE: R IX N  
 أَخْر E AE: KH AA R  
 أَخْر E AA: KH AA R  
 أَخْرَ E AE: KH AA R AA  
 أَخْرُونْ E AE: KH AA R UW N AE

أَخْرَيْنَ E AE: KH AA R IX: N AE  
 أَخْرَيْنُ E AE: KH AA R IX: N  
 أَخَذَ E AE: KH IX DH AE T UH N  
 آذَارُ E AE: DH AE: R  
 أَرُ E AE: R  
 أَسْبَا E AE: S Y AE:  
 أَسْبَانُ E AE: S Y AE: N  
 أَسْبَوِيَّةُ E AE: S Y AE W IH Y AE H  
 2) أَسْبَوِيَّةُ E AE: S Y AE W IH Y AE T  
 أَفَاقُ E AE: F AE: Q IX  
 أَفَاقُ E AE: F AE: Q  
 أَفَاقُ E AE: F AA: Q  
 أَفَاقُ E AA:: F AE: Q  
 آل E AE: L  
 آلَافُ E AE: L AE: F IH N  
 آلَافُ E AE: L AE: F  
 آلَافُ E AE: L AE: F IH

#### 5 Evaluation of the AASR System

For large vocabulary systems, the performance of the Sphinx system was tested extensively on the DARPA Hub-4 Broadcast news project [22]. The HUB-4 Broadcast News Speech Corpus contains a total of 104 hours of broadcasts from various television networks and radio networks with the corresponding transcripts. The English evaluation test material (1.5 hours) is administered by the NIST. The acoustic models used for this test had 5000 tied states with 32 Gaussians per state. A trigram LM with 4.7M bigrams and 15.5M trigrams covering a vocabulary of 64,000 words was used. The following table gives a summary of the sphinx performance.

Vocabulary Size	Test	% WER
11	TIDIGITS	0.661
79	AN4	1.300
1,000	RM1	2.746
5,000	WSJ5K	7.323
60,000	HUB4	18.845

Table 3: Performance of the Sphinx engine under various speech recognition tasks\*.

\*<http://cmusphinx.sourceforge.net/sphinx4/>

The developed AASR was based on 5.4 hour of Arabic broadcast news. 4.3 hours are used in training, and the remaining (20%) is used for testing. The corpus vocabulary came to 14232 words. The state distributions of all triphones are tied into 1132 senons, each is represented by a Gaussian mixtures of 8 components. The number of test utterances was 1144, consisting of a total of 9288 words. Word Error Rate (WER) was initially 12.55 %. 8318 words were correctly recognized. The analysis of the error indicates that there was 882 word substitution errors, 196 word insertion errors, and 88 word deletion errors.

Following this initial results, extensive testing and tuning of some recognition parameters were carried out. It was found significant improvement can be achieved by accounting for the noise, which causes a large number of insertion errors. The filler dictionary was extended by adding inhalation noise and background noise. Although the models for these noise phones were trained based on a limited number of utterances, the impact was significant.

The correctly recognized words was 90.13%, and WER came down to 11.3 %. Word substitution errors drops from 882 to 832 words, word insertion errors dropped also from 196 cases to 171 cases. The word deletion cases were 85 cases.

Further analysis indicates that many of the word substitution errors are due to slight differences (deletion/substitution) of diacritical marks only. Since MSA text is written without diacritical marks, the error analysis was carried out once more after removing all the diacritical marks. The percent of the correctly recognized words was 92.84%. The WER dropped to 9.0%. The number of word substitution cases was reduced to 580 cases. However, the number of word deletion and word insertion errors did not change. The following table, Table 4, shows a sample of the recognition result, where the right column is the recognition result, and the left column is the original test text.

Original text	Recognition result
وممثلين عن عدد من الدول الأوروبية الخاص بتحبيد أسهمها	وممثلين عن إن الدول الأوروبية الخاص بتحبيد أسهمها
ويعد القرض الأحدث ضمن ثلاثة مليارات ونصف المليار دولار	ويعد القرض الأحدث ضمن ثلاثة مليارات ونصف المليار دولار
سيتقابلان وجها لوجه في المباراة النهائية	سيتقابلان وجها لوجه المباراة النهائية
وتستهدف خطة القطاع خلال العام المقبل رفع قدرات توليد الطاقة الكهربائية لتصل إلى ألف ميقات	وتستهدف خطة القطاع خلال العام المقبل رفع قدرات توليد الطاقة الكهربائية لتصل إلى أي في مليارات
متجاوزا مليار دولار للعام الثاني على التوالي	متجاوزا مليار دولار للعام الثاني على التوالي
حسب استفتاء جماهيري أجرته إذاعة بي بي سي البريطانية	حسب استفتاء جماهيري أجرته إذاعة أيبب بالسلع البريطانية
من خلال بيعه في صفقة عاجلة لنادي سي ميلانو الإيطالي	من خلال بيعه في صفقة عاجلة الفارس والمدرب
في الدوري الدرجة الأولى الليبي لكرة القدم لهذا الموسم	الدور الدرجة الأولى الليبي لكرة القدم لهذا الموسم
مدعومة بأبناء طيبة من شركات كبرى	مدعومة بأبناء طيبة من شركات كبرى

Table 4: Sample of the unvoiced recognition results.

In HMM-based ASR systems higher recognition accuracies generally result from detailed models with a large number of parameters, and by a detailed search of all possible hypotheses during recognition. Since computation time increases with the number of parameters in the acoustic models, and also with the number of hypotheses considered during recognition, the requirement of high recognition accuracy is a trade-off with the requirement of high

recognition speed. The compromise in accuracy involved in achieving higher speeds is lower if a large amount of system memory is available to store intermediate results during recognition, which can be used to achieve higher accuracies. The model was built using 8 Gaussian mixtures. Using 16 Gaussian mixtures will be considered at later stage, however, for these models to be properly trained larger corpus needs to be developed as well.

## Conclusion

The paper reports the progress in an on-going research towards achieving large vocabulary, speaker independent, natural Arabic automatic speech recognition system. During this initial phase an infrastructure for research was developed, and a 5.4 hour corpus was built. A rule-based phonetic dictionary was also developed. The speech recognition system achieves a comparable accuracy to English ASR system for the same vocabulary size. Further enhancement will be carried out during the next phase of this research work, including extending the corpus to 40 hours, enhancing the rule based phonetic dictionary, and using a finer parameterization of the acoustic model.

## Acknowledgments

This work was supported by a grant #AT-24-94 by King Abdulaziz City of Science and Technology. The authors would like also to thank King Fahd University of Petroleum and Minerals for its support in carrying out this project.

## References

- [1] Alghamdi, Mansour , Arabic Phonetics, Attaobah, Riyadh, 2000.
- [2] Algamdi, Mansour, KACST Arabic Phonetics Database, *The Fifteenth International Congress of Phonetics Science*, Barcelona, 3109-3112, 2003.
- [3] Alghamdi, Mansour, Mustafa Elshafei and Husni Almuhtasib, Speech Units for Arabic Text-to-speech, *The Fourth Workshop on Computer and Inforamtion Sciences*, 199-212, 2002.
- [4] A.M. Alimi, M. Ben Jemaa , "Beta Fuzzy Neural Network Application in Recognition of Spoken Isolated Arabic Words", *International Journal of Control and Intelligent Systems*, Special Issue on Speech Processing Techniques and Applications, Vol. 30, No.2 , 2002.
- [5] Bahi, H.; Sellami, M. " A hybrid approach for Arabic speech recognition", *ACS/IEEE International Conference on Computer Systems and Applications*, 2003. 14-18 July 2003,
- [6] Billa, J.; Noamany, M.; Srivastava, A.; Liu, D.; Stone, R.; Xu, J.; Makhoul, J.; Kubala, F.," Audio indexing of Arabic broadcast news", *Proceedings. (ICASSP '02). IEEE International Conference on Acoustics, Speech,*

- and *Signal Processing*, 2002. Volume 1, 2002 Page(s):I-5 - I-8 vol.1
- [7] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-cambridge toolkit," in *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, Sept. 1997.
- [8] El Choubassi, M.M.; El Khoury, H.E.; Alagha, C.E.J.; Skaf, J.A.; Al-Alaoui, M.A. "Arabic speech recognition using recurrent neural networks", *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology*, 2003. ISSPIT pp. 543- 547 , Dec. 2003.
- [9] El-Ramly, S.H.; Abdel-Kader, N.S.; El-Adawi, R, "Neural networks used for speech recognition", *Radio Science Conference, 2002. (NRSC 2002). Proceedings of the Nineteenth National* , pp. 200-207, March 2002.
- [10] M. Elshafei Ahmed, " Toward an Arabic Text-to-Speech System", *The Arabian Journal of Science and Engineering*, Vol. 16, No. 4B, pp.565-583, 1991.
- [11] Moustafa Elshafei, Husni Almuhtasib and Mansour Alghamdi, Techniques for High Quality Text-to-speech, *Information Science*, 140 (3-4) 255-267, 2002.
- [12] Moustafa Elshafei, Husni Al-Muhtaseb and Mansour Alghamdi, "Statistical Methods for Automatic Diacritization of Arabic text", *Proceedings 18th National computer Conference NCC' 18*, Riyadh, March 26-29, 2006.
- [13] Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Alghamdi, "Machine Generation of Arabic Diacritical Marks", *Proceedings of the 2006 International Conference on Machine Learning; Models, Technologies, and Applications (MLMTA'06)*, June 2006, USA.
- [14] Hussein A.R. Hiyassat, Automatic Pronunciation Dictionary Toolkit for Arabic Speech Recognition Using SPHINX Engine, Ph.D., Arab Academy for Banking and Financial Sciences, Amman, Jordan, 2007.
- [15] HTK speech recognition tool kit.  
<http://htk.eng.cam.ac.uk/>
- [16] X. Huang, F. Alleva, H. W. Hon, M. Y. Hwang, and R. Rosenfeld, "The SPHINX-II speech recognition system: an overview," *Computer Speech and Language*, vol. 7, no. 2, pp. 137–148, 1993.
- [17] X.Huang, A. Acero, and H. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.
- [18] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1998.
- [19] P. Lamere, P. Kwok, W. Walker, E. Gouvea, R. Singh, B. Raj, and P. Wolf, "Design of the CMU Sphinx-4 decoder," in *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 1181–1184
- [20] K.F. Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System," PhD Thesis, *Carnegie Mellon University*, 1988.
- [21] Fahad A.H. Al-Otaibi, Speaker-Dependant Continuous Arabic Speech Recognition, M.Sc. Thesis, *King Saud University*, 2001.
- [22] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer, "The 1996 HUB-4 Sphinx-3 system," in *Proceedings of the DARPA Speech Recognition Workshop*. Chantilly, VA: DARPA, Feb. 1997.  
[http://www.nist.gov/speech/publications/darpa97/pdf/pla\\_cewa1.pdf](http://www.nist.gov/speech/publications/darpa97/pdf/pla_cewa1.pdf)
- [23] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1988, pp. 651–654.
- [24] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77(2), 1989.
- [25] L. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
- [26] Shoaib, M.; Rasheed, F.; Akhtar, J.; Awais, M.; Masud, S.; Shamail, S., "A novel approach to increase the robustness of speaker independent Arabic speech recognition", *7th International Multi Topic Conference, 2003. INMIC 2003*. 8-9 Dec. 2003, pp. 371- 376.
- [27] Sphinx-4 Java-based Speech Recognition Engine,  
<http://cmusphinx.sourceforge.net/sphinx4/>
- [28] Young, S. (1996), "A review of large-vocabulary continuous-speech recognition", *IEEE Signal Processing Magazine*, pages 45-57, 2007.