Automated Speech Recognition System (ASR)

Mohamed Belgacem[1], **Georges Antoniadis**[2], **and Mounir Zrigui**[3]
[1]Laboratoire LIDILEM , University Grenoble, FRANCE
[2]Laboratoire LIDILEM , University Grenoble, FRANCE
[3]Laboratoire UTIC, University Monastir, TUNISIE

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

1 Introduction

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone. It has a wide area of applications: command recognition (voice user interface with the computer), dictation, interactive voice response, it can be used to learn a foreign language. ASR can help also, handicapped people to interact with society. It is a technology which makes life easier and very promising [8].

View the importance of ASR too many systems are developed, the most popular are: Dragon Naturally Speaking, IBM via voice, Microsoft SAPI. Open source speech recognition systems are available too, such as [24, 5, 17, 16, 11, 19]. We are interested in exploring this last, which is based on Hidden Markov Models (HMMs) [8]. A Hidden Markov Model (HMM) is a statistical model where the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters, from the observable parameters, based on this assumption. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. Its extension into foreign languages (English is the standard) represent a real research challenge area.

Although Arabic is currently one of the most widely spoken language in the world, there has been relatively little speech recognition research on Arabic compared to the other languages [14, 23, 22]. The first works on Arabic ASR has concentrated on developing recognizers for Modern Standard Arabic (MSA). The most difficult problems in developing highly accurate ASRs for Arabic are the predominance of non diacritized text material, the enormous dialectal variety, and the morphological complexity.

Kirchhoff et al. [23] investigate the recognition of dialectal Arabic and study the discrepancies between dialectal and formal Arabic in the speech recognition point of view. Vergyri et al. [22] investigate the use of morphology-based language model at different stages in a speech recognition system for conversational Arabic; they studied also the automatic diacritizing Arabic text for use in acoustic model training for ASR. In their previous papers Satori et al. [21, 20], introduce an Arabic voice recognition system where both training and recognizing process use Romanized characters.

Most of previous works on Arabic ASR have been concentrated on developing recognizers using Romanized characters. In this work we investigate a system using entirely Arabic environment based on the HMM statistical approach and 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.

We have generated a pronunciation dictionary and trained acoustic model with Arabic speech data. In the next section we present a brief description of the Arabic language. In section 3, we describe the Arabic speech recognition system and our investigations to adapt the system to Arabic language. In section 4, we present experimental results.

Finally, in section 5, we provide our conclusions and future directions.

2 Arabic languages

Arabic is a Central Semitic language, thus related to and classified alongside other Semitic languages such as Hebrew and the Neo-Aramaic languages. Arabic has more speakers than any other language in the Semitic language family. It is spoken by more than 280 million people as a first language, most of whom live in the Middle East and North Africa. It is the official language of 22 countries and it is the liturgical language of Islam since it is the language of the Quran, the Islamic Holy Book. Arabic has many different, geographically distributed spoken varieties, some of which are mutually unintelligible. Modern Standard Arabic (sometimes called Literary Arabic) is widely taught in schools, universities, and used in workplaces, government and the media.

Table 1 shows the listing of the phoneme set used in the training and the corresponding phoneme symbols. The table also shows illustrative examples of the vowel usage. The phoneme set chosen is based on the previous experience with Arabic text-to-Speech systems [4, 5, and 10], and the corresponding phoneme set which is successfully used in the English ASR [3].

Table 1. The complete phoneme set used in training

Phoneme	Letter	Phoneme	Letter
/AE/	◄َ (ahtaF) ب	/KH/	(hahK) خ
/AE:/	بأب ◄ا	/D/	(laD) د
/AA/	خَ◄َ	/DH/	(lahT) ذ
/AH/	دَق◄َ	/R/	(heR) ر
/UH/	بُ◄ُ (Damma)	/Z/	(niaZ) ز
/UW/	نوُد ◄ُ و	/S/	(neeS) س
/UX/	نصُغ ◄ُ	/SH/	(neehS) ش
/IH/	بـ تنِ ◄ِ (Kasra)	/SS/	(daS) ص
/IY/	ليف ◄ي	/DD/	daD) ض)
/IX/	فنِص ◄ِ	/TT/	(haT) ط
/AW/	موَل ◄وَ	/DH2/	(hahT) ظ
/AE/	◄َ (ahtaF) ب	/KH/	(hahK) خ
/AE:/	بأب ◄ا	/D/	(laD) د
/AA/	خَ◄َ	/DH/	(lahT) ذ
/AH/	دَق◄َ	/R/	(heR) ر
/AY/	فيَض ◄يَ	/AI/	(niA) ع
/UN/	يجنُن ◄ُ	/GH/	(niahG) غ
/AN/	مَن◄َ	/F/	(heF) ف
/IN/	امِم ◄ِ	/V/	ف◄ ازيِ ف
/E/	(azmaH) ء	/Q/	(faQ) ق
/B/	(heB) ب	/K/	(faK) ك
/T/	(heT) ت	/L/	(maL) ل
/TH/	(hehT) ث	/M/	(meeM) م

The regular Arabic short vowels are /AE/. /IH/, and /UH/ corresponding to the Arabic diacritical marks Fatha, Damma, and Kasra respectively. The /AA/ is the pharyngealized allophone of /AE/, which appears after an emphatic letter. Similarly, the /IX/ and /UX/ are the pharyngealized allophones of /IH/ and /UH/ respectively. When /AE/ appears before an emphatic letter, its allophone /AH/ is used instead. When a short vowel is located between two nasal leters in the same syllable it is likely to be nasalized. The allophones /AN/, /IN/, and /UN/ are the nasalized versions of /AE/, /IH/, and /UH/ respectively.

The regular Arabic long vowel allophones are /AE:/ /IY/ and /UW/ respectively. The length of a long vowel is normally equal to two short vowels. The allophones /AY/ and /AW/ are actually two vowel sounds in which the articulators move from one post to another. These vowels are called Diphthongs. The allophone /AY/ appears when a Fatha comes before an undiacritized Yeh. Similarly, /AW/ appears when a Fatha comes before an undiacritized Waw.

The Arabic voiced stops phonemes /B/ and /D/ are similar to their English counter parts. /DD/ corresponds to the sound of the Arabic Dhad letter.

The Arabic voiceless stops /T/ and /K/ are basically similar to their English counter parts.

The sound of the Arabic emphatic letter Qaf is represented by the phone /Q/. The Hamza plosive sound is represented by the phone /E/, and the sound of Jeem (in many dialects) is represented by /G/.

The voiceless fricatives are produced with no vibration of the voice cords. The sound is produced by the turbulence flow of air through a constriction. The Arabic voiceless fricatives /F/, /S/, /TH/, /SH/, and /H/ are basically similar to their English twins. In addition, the Arabic phones /SS/, /HH/, and /KH/ are the sounds of the Arabic letters Sad, Hah, and Khah respectively.

Voiced Fricatives are generated with simultaneous vibration of the vocal cords. The Arabic voiced fricative phones are /AI/, /GH/, /Z/, and /DH/ corresponding to the sound of the Arabic letters: Ain, Ghain, Zain, and Thal.

The Arabic affricative sound /JH/ is similar to the corresponding one in English, while /ZH/ is a concatenation of a voiced stop followed by a fricative sound.

The Arabic resonants are similar to the the English resonant phones. These are /Y/ for Yeh, /W/ for Waw, /L/ for Lam, and /R/ for Reh.

3 Trainings steps

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

3.1 Acoustic model training

The training procedure consists of three phases. 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 contextdependant (CD) phase, triphones are added to the HMM set. In the model definition stage, all the possible triphones will be created, and then the tirphones 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 reestimation is performed iteratively.

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 reestimated 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.

4 Arabic broadcast news corpus

The audio files were recorded from several TV news channels at a sampling rate of 20 ksps. A total of 468 news stories, summing up to 13 hours of speech, were recorded and split into 8379 files with an average file length of 7 seconds. The length of wave files range from 0.8 seconds to 18 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.

5 Transcription

All the 8379 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.

To speed up the diacritization, an phonetiser of Arabic was developed by the authors for automatic vocalization of the Arabic text. 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 94% since, the system was trained on different text subjects. Hand editing was then necessary to bring the transcription to the desired accuracy level.

6 Arabic phonetic dictionary

Using the selected phoneme set, we developed 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 witch 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 (ٱ) 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 34894 entries. A sample from the developed phoneme dictionary is listed below.

Table2. A sample from the developed phoneme dictionary

أمس	S M AE E
الدائرة	H AE R IH E AE: D EL
الجنائية	H AE Y IH E AE : N IH J EL
الرابعة	H AE AI IH B AE: R EL
بالمحكمة	H AE M AE K AE M EL IH B
الابتدائية	H AE Y IH E AE : D IH T B IH E EL
بتونس	S IH N UW T IH B
النظر	AE R AE DH2 AE N EL
آسْيَويَّة	H AE Y IH W AE Y S AE: E
2 آسْيَويَّة	T AE Y IH W AE Y S AE: E

7 Evaluation of AASR system

The developed AASR was based on 13 hours of Arabic broadcast news. 10.4 hours are used in training, and the remaining (20%) is used for testing. The corpus vocabulary came to 28783 words. The number of test utterances was 2304, consisting of a total of 17928 words. Word Error Rate (WER) was initially 9.55 %. 16215 words were correctly recognized. The analysis of the error indicates that there was 1712 word substitution errors, 1220 word insertion errors, and 492 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 correctly recognized word was 93%. The analysis of the error indicates that there was 1712 to 1254 word substitution errors, 960 word insertion errors, and 294 word deletion errors.

8 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 11 hours 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 100 hours, enhancing the rule based phonetic dictionary, and using a finer parameterization of the acoustic model.

9 References

[1] Alghamdi, Mansour , Arabic Phonetics, Attaoobah, 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 Commu nication 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 Pronuncation 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, Geneve, 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/placewa1.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.