Speech Recognition System of Arabic Digits based on A Telephony Arabic Corpus

Yousef Ajami Alotaibi[1], **Mansour Alghamdi**[2], **Fahad Alotaiby**[3]
[1]Computer Engineering Department, King Saud University, Riyadh, Saudi Arabia
[2]King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia
[3]Department of Electrical Engineering, King Saud University, Riyadh, Saudi Arabia

Abstract - *Automatic recognition of spoken digits is one of the difficult tasks in the field of computer speech recognition. Spoken digits recognition process is required in many applications such as speech based telephone dialing, airline reservation, automatic directory to retrieve or send information, etc. These applications take numbers and alphabets as input. Arabic language is a Semitic language that differs from European languages such as English. One of these differences is how to pronounce the ten digits, zero through nine. In this research, spoken Arabic digits are investigated from the speech recognition problem point of view. The system is designed to recognize an isolated whole-word speech. The Hidden Markov Model Toolkit (HTK) is used to implement the isolated word recognizer with phoneme based HMM models. In the training and testing phase of this system, isolated digits data sets are taken from the telephony Arabic speech corpus, SAAVB. This standard corpus was developed by KACST and it is classified as a noisy speech database. A hidden Markov model based speech recognition system was designed and tested with automatic Arabic digits recognition. This recognition system achieved 93.72% overall correct rate of digit recognition.*

Keywords: Arabic, digits, SAAVB, HMM, Recognition, Telephony corpus.

1 Introduction

1.1 Arabic Language

Arabic is a Semitic language, and it is one of the oldest languages in the world. Currently it is the second language in terms of number of speakers [1]. Arabic is the first language in the Arab world, i.e., Saudi Arabia, Jordan, Oman, Yemen, Egypt, Syria, Lebanon, etc. Arabic alphabets are used in several languages, such as Persian and Urdu. Standard Arabic has basically 34 phonemes, of which six are vowels, and 28 are consonants [2]. A phoneme is the smallest element of speech units that indicates a difference in meaning, word, or sentence. Arabic language has fewer vowels than English language. It has three long and three short vowels, while American English has twelve vowels [3].

Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. These two classes can be found only in Semitic languages like Hebrew [2], [4]. Arabic digits zero to nine (sĕfr, wâ-hĕd, ‘aâth-nāyn, thâ-lă-thâh, ‘aâr-bâ-‘aâh, khâm-sâh, sĕt-tâh, sûb-‘aâh, thâ-mă-nĕ-yĕh, and tĕs-âh) are polysyllabic words except the first one, zero, which is a monosyllable word [2]. The allowed syllables in Arabic language are: CV, CVC, and CVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant [2]. Table 1 shows the ten Arabic digits along with the way of how to pronounce them in Modern Standard Arabic (MSA), number and types of syllables in every spoken digit.

Table 1: Arabic Digits

Digit	Arabic Writing	Pronunciation	Syllables
1	واحد	wâ-hĕd	CV-CVC
2	أثنين	‘aâth-nā-yn	CVC-CVCC
3	ثلاثة	thâ-lă-thâh	CV-CV-CVC
4	أربعة	‘aâr-bâ-‘aâh	CVC-CV-CVC
5	خمسة	khâm-sâh	CVC-CVC
6	ستة	sĕt-tâh	CVC-CVC
7	سبعة	sûb-‘aâh	CVC-CVC
8	ثمانية	thâ-mă-nĕ-yĕh	CV-CV-CV-CVC
9	تسعة	tĕs-âh	CVC-CVC
0	صفر	sĕfr	CVCC

1.2 Spoken Alpha digits Recognition

In general, spoken alphabets and digits for different languages were targeted by automatic speech recognition researchers. A speaker-independent spoken English alphabet recognition system was designed by Cole et al. [5]. That system was trained on one token of each letter from 120 speakers. Performance was 95% when tested on a new set of 30 speakers, but it was increased to 96% when tested on a second token of each letter from the original 120 speakers.

Other efforts for spoken English alphabets recognition was conducted by Loizou et al. [6]. In their system a high performance spoken English recognizer was implemented using context-dependent phoneme hidden Morkov models (HMM). That system incorporated approaches to tackle the problems associated the confusions occurring between the stop consonants in the E-set and the confusions between the nasal sounds. That recognizer achieved 55% accuracy in

nasal discrimination, 97.3% accuracy in speaker-independent alphabet recognition, 95% accuracy in speaker-independent E-set recognition, and 91.7% accuracy in 300 last names recognition.

Karnjanadecha et al. [7] designed a high performance isolated English alphabet recognition system. The best accuracy achieved by their system for speaker independent alphabet recognition was 97.9%. Regarding digits recognitions, Cosi et al. [8] designed and tested a high performance telephone bandwidth speaker-independent continuous digit recognizer. That system was based on artificial neural network and it gave a 99.92% word recognition accuracy and 92.62% sentence recognition accuracy.

Arabic language had limited number of research efforts compared to other languages such as English and Japanese. A few researches have been conducted on the Arabic alphadigits recognition. In 1985, Hagos [9] and Abdullah [10] separately reported Arabic digit recognizers. Hagos designed a speaker-independent Arabic digits recognizer that used template matching for input utterances. His system is based on the LPC parameters for feature extraction and log likelihood ratio for similarity measurements. Abdullah developed another Arabic digits recognizer that used positive-slope and zero-crossing duration as the feature extraction algorithm. He reported 97% accuracy rate. Both systems mentioned above are isolated-word recognizers in which template matching was used. Al-Otaibi [11] developed an automatic Arabic vowel recognition system. Isolated Arabic vowels and isolated Arabic word recognition systems were implemented. He studied the syllabic nature of the Arabic language in terms of syllable types, syllable structures, and primary stress rules.

1.3 Hidden Markov Models and Used Tools

Automatic Speech Recognition (ASR) systems based on the HMM started to gain popularity in the mid-1980's [6]. HMM is a well-known and widely used statistical method for characterizing the spectral features of speech frame. The underlying assumption of the HMM is that the speech signal can be well characterized as a parametric random access, and the parameters of the stochastic process can be predicted in a precise, and well-defined manner. The HMM method provides a natural and highly reliable way of recognizing speech for a wide range of applications [12], [13].

The Hidden Markov Model Toolkit (HTK) [14] is a portable toolkit for building and manipulating HMM models. It is mainly used for designing, testing, and implementing ASR systems and related research tasks. This research concentrated on analysis and investigation of the Arabic digits from an ASR perspective. The aim is to design a recognition system by using the Saudi Accented Arabic Voice Bank (SAAVB) corpus provided by King Abdulaziz City for Science and Technology (KACST). SAAVB is considered as a noisy speech database because most of the part of it was recorded in normal life conditions by using mobile and other telephone lines [15]. The system is based on HMMs and with the aid of HTK tools.

2 Experimental Framework

2.1 System Overview

A complete ASR system based on HMM was developed to carry out the goals of this research. This system was divided into three modules according to their role. The first module is training module, whose function is to create the knowledge about the speech and language to be used in the system. The second subsystem is the HMM models bank, whose function is to store and organize the system knowledge gained by the first module. Final module is the recognition module, whose function is tried to figure out the meaning of the input speech given in the testing phase. This is done with the aid of the HMM models mentioned above.

As illustrated in Table 2, the parameters of the system were 8KHz sampling rate with a 16 bit sample resolution, 25 millisecond Hamming window duration with a step size of 10 milliseconds, MFCC coefficients with 22 as the length of cepstral leftering and 26 filter bank channels of which 12 were as the number of MFCC coefficients, and of which 0.97 were as the pre-emphasis coefficients.

Table 2: System parameters

Parameter	Value
Sampling rate	8KHz, 16 bits
Database	Isolated 10 Arabic digits
Speakers	1033 male and female
Condition of Noise	Normal life
Accent	Saudi (from different regions)
Preemphased	$1-0.97z^{-1}$
Window type	Hamming, 25 mseconds
Window step size	10 mseconds
SAAVB Parts	SAAVB-01, -02, -03

Phoneme based models are good at capturing phonetic details. Also context-dependent phoneme models can be used to characterize formant transition information, which is very important to discriminate between digits that can be confused. The Hidden Markov Model Toolkit (HTK) is used for designing and testing the speech recognition systems throughout all experiments. The baseline system was initially designed as a phoneme level recognizer with three active states, one Gaussian mixture per state, continuous, left-to-right, and no skip HMM models. The system was designed by considering all thirty-four Modern Standard Arabic (MSA) monophones as given by the KACST labeling scheme given in [16]. This scheme was used in order to standardize the phoneme symbols in the researches regarding classical and MSA language and all of its

variations and dialects. In that scheme, labeling symbols are able to cover all the Quranic sounds and its phonological variations. The silence (sil) model is also included in the model set. In a later step, the short pause (sp) was created from and tied to the silence model. Since most of the digits are consisted of more than two phonemes, context-dependent triphone models were created from the monophone models mentioned above. Before this, the monophone models were initialized and trained by the training data explained above. This was done by more than one iteration and repeated again for triphones models. A decision tree method is used to align and tie the model before the last step of training phase. The last step in the training phase is to re-estimate HMM parameters using Baum-Welch algorithm [12] three times.

2.2 Database

The SAAVB corpus [15] was created by KACST and it contains a database of speech waves and their transcriptions of 1033 speakers covering all the regions in Saudi Arabia with statistical distribution of region, age, gender and telephones. The SAAVB was designed to be rich in terms of its speech sound content and speaker diversity within Saudi Arabia. It was designed to train and test automatic speech recognition engines and to be used in speaker, gender, accent, and language identification systems. The database has more than 300,000 electronic files. Some of those files contained in SAAVB are 60,947 PCM files of recorded speech via telephone, 60,947 text files of the original text, 60,947 text files of the speech transcription, and 1033 text files about the speakers. The mentioned files have been verified by IBM Egypt and it is completely owned by KACST.

The used parts of SAAVB corpus in this research are those parts that contain digits only. Depending on SAAVB datasheet [15] parts that contain Arabic digits are SAAVB-1 (one-digit per recording selected randomly from the ten Arabic digits), SAAVB-2 (ten-digit number per recording), SAAVB-3 (five-digit number per recording). These three parts contain more than 15,480 recorded digits. These three subsets are those parts dealing with uttering isolated Arabic digits. We partitioned these parts of corpus (i.e., digits part) into two disjoint sets, the first one for the training which constitutes 75% of digits parts (i.e., SAAVB-1, SAAVB-2, and SAAVB-3) and contain more than 11,600 recorded digits. On the other hand, the second one is for testing phase which contains 25% and contains more than 3,850 recorded digits.

3 Results

Table 3 shows the correct rate of the system for digits individually in addition to the system overall correct rate. Depending on testing database set, the system must try to recognize 3,870 samples for all 10 digits. The overall system performance was 93.72%, which is reasonably high where our database is considered as noisy corpus. Our initial audio files analysis showed that Signal to Noise Ratio (SNR) is ranging from 6dB (with noisy environment) to 34dB (with relatively less noisy environment). The system failed in recognizing 275 (243 were missed plus 32 were deleted) digits out of 3,870 recorded digits as shown in Table 3. Digits 1, 2, 3, and 8 have gotten reasonably high correct rate of recognition; on the other hand, the worst performance was encountered with digits 0 where the performance was equal to only 82.82% where the system failed in recognizing 75 (67 were missed plus 8 were deleted) digits out 398 tokens were miss-recognized depending on figures given Table 3. Even though the database size is medium (only the ten spoken Arabic digits) and with the existence of noise, the system showed an unexpected high performance due to the variability in how to pronounce Arabic digits.

Table 3: Confusion matrix

	zero	one	two	three	four	five	six	seven	eight	nine	Del	Acc. (%)	Tokens	Missed
zero	323	0	11	25	3	2	11	6	3	6	8	82.82	390	67
one	2	381	0	5	1	1	1	1	0	0	4	97.19	392	11
two	4	1	386	4	0	0	0	1	0	2	1	96.98	398	12
three	3	1	0	378	2	1	1	1	1	2	1	96.92	390	12
four	4	1	3	9	386	0	0	9	1	1	3	93.24	414	28
five	2	1	1	2	2	326	4	3	0	2	2	95.04	343	17
six	7	2	4	9	0	0	352	0	2	9	3	91.43	385	33
seven	4	0	1	8	1	2	0	367	2	0	3	95.32	385	18
eight	1	0	1	7	0	0	2	0	351	0	6	96.96	362	11
nine	5	1	4	11	1	1	8	2	1	345	1	91.03	379	34
Ins	374	33	173	1099	23	12	123	38	26	55				
Total											32	93.72	3870	243

Table 4: Digits that were picked in case of miss-recognition for all digits

Digit	Mostly Confused with
0	2, 3, 6, 7, 9
1	3
2	3
3	0
4	3, 7
5	6
6	0, 3, 9
7	3
8	3
9	0, 3, 6

From Tables 3 and 4, it can be seen that digit 3 is picked in case of missing the correct digit for most of digits. This happened in analyzing errors of digits 0, 1, 2, 4, 6, 7, 8, and 9. We think that this is due to the similarity of many parts of digit 3 to noise such as phoneme /θ/ which exist twice in that digit. Also we can notice that digit 3 got one of the highest correct rate among Arabic digits. The investigation of errors of the recognition system is one of the scheduled tasks in this project, and we are going to do it in the near future. All digits appeared in the test data subset at least 345 times for all speakers, genders, ages, and regions in Saudi Arabia. These results are initial and the research is going to continue to analyze all subsets of the corpus, noise level and

effect, different parameters of the system such as increasing number of mixtures per state, and the effect of similarity among the ten Arabic digits.

4 Conclusion

To conclude, a spoken Arabic digits recognizer is designed to investigate the process of automatic digits recognition. This system is based on HMM and by using Saudi accented and noisy corpus called SAAVB. This system is based on HMM strategy carried out by HTK tools. The system consists of training module, HMM modules storage, and recognition module. The SAAVB corpus supplied by KACST is used in this research. Arabic language is investigated in this research in addition to the evaluation of SAAVB corpus. The overall system performance is 93.72%. The best correct rates were encountered in the case of more than one digit, but the worst correct rate was encountered in the case of digit 0. These results are our initial ones and the work is going to continue to cover all aspect of the system including: performance, corpus, noise effect, Arabic digits, and the suggestion regarding this point of research.

5 Acknowledgment

This paper is supported by KACST (Project: 28-157).

6 References

[1] MSN encarta. “Languages Spoken by More Than 10 Million People”; http://encarta.msn.com/media_701500404/ Languages_Spoken_by_More_Than_10_Million_People.html, 2007.

[2] Muhammad Alkhouli. “Alaswaat Alaghawaiyah”; Daar Alfalah, Jordan, 1990 (in Arabic).

[3] J. Deller, J.Proakis, and J. H.Hansen. “Discrete-Time Processing of Speech Signal”; Macmillan, 1993.

[4] M. Elshafei. “Toward an Arabic Text-to-Speech System”; The Arabian Journal for Scince and Engineering, Vol. No. 16, Issue No. 4B, pp. 565-83, Oct. 1991.

[5] R. Cole, M. Fanty, Y. Muthusamy, and M. Gopalakrishnan. “Speaker-Independent Recognition of Spoken English Letters”; International Joint Conference on Neural Networks (IJCNN), Vol. No. 2, pp. 45-51, Jun.1990.

[6] P. C. Loizou and A. S. Spanias. “High-Performance Alphabet Recognition”; IEEE Trans. on Speech and Audio Processing, Vol. No. 4, Issue No. 6, pp. 430-445, Nov. 1996.

[7] M. Karnjanadecha and Z. Zahorian. “Signal Modeling for High-Performance Robust Isolated Word Recognition”; IEEE Trans. on Speech and Audio Processing, Vol. No. 9, Issue No. 6, pp. 647-654, Sep. 2001.

[8] P. Cosi, J. Hosom, and A. Valente. “High Performance Telephone Bandwidth Speaker Independent Continuous Digit Recognition”; Automatic Speech Recognition and Understanding Workshop (ASRU), Trento, Italy, 2001.

[9] Elias Hagos. “Implementation of an Isolated Word Recognition System”; UMI Dissertation Service, 1985.

[10] W. Abdulah and M. Abdul-Karim. “Real-time Spoken Arabic Recognizer”; Int. J. Electronics, Vol. No. 59, Issue No. 5, pp.645-648, 1984.

[11] A. Al-Otaibi. “Speech Processing”; The British Library in Association with UMI, 1988.

[12] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”; Proceedings of the IEEE, Vol. No. 77, Issue No. 2, pp. 257-286, Feb. 1989.

[13] B. Juang and L. Rabiner. “Hidden Markov Models for Speech Recognition”; Technometrics, Vol. No. 33, Issue No. 3, pp. 251-272, Aug. 1991.

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. “The HTK Book (for HTK Version. 3.4)”; Cambridge University Engineering Department, http:///htk.eng.cam.ac.uk/prot-doc/ktkbook.pdf, 2006.

[15] M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairi, and M. Aldusuqi. “Saudi Accented Arabic Voice Bank (SAAVB)”; Final report, Computer and Electronics Research Institute, King Abdulaziz City for Science and technology, Riyadh, Saudi Arabia, 2003.

[16] Mansour Alghamdi, Yahia El Hadj, and Mohamed Alkanhal. “A Manual System to Segment and Transcribe Arabic Speech”; IEEE International Conference on Signal Processing and Communication (ICSPC07), Dubai, UAE, 24–27 Nov. 2007.