

Speech Recognition System of Arabic Alphabet based on a Telephony Arabic Corpus

Yousef Ajami Alotaibi¹, Mansour Alghamdi², Fahad Alotaiby³

¹Computer Engineering Department, King Saud University, Riyadh, Saudi Arabia

²King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

³Department of Electrical Engineering, King Saud University, Riyadh, Saudi Arabia
yaalotaibi@ksu.edu.sa, mghamdi@mghamdi.com, falotaiby@hotmail.com

Abstract - Automatic recognition of spoken alphabets is one of the difficult tasks in the field of computer speech recognition. In this research, spoken Arabic alphabets 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 alphabets 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 alphabets recognition. Four different experiments were conducted on these subsets, the first three trained and tested by using each individual subset, the fourth one conducted on these three subsets collectively. The recognition system achieved 64.06% overall correct alphabets recognition using mixed training and testing subsets collectively.

Keywords: Arabic, alphabets, 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 fifth language in terms of number of speakers [1]. Arabic is the native language of twenty-five countries including Saudi Arabia, Jordan, Oman, Yemen, Egypt, Syria, Lebanon, etc [1]. Arabic alphabets are used in several languages in addition to Arabic, 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]. The allowed syllables in Arabic language are: CV, CVC, and CvCC where V indicates long/short vowels, v indicates short vowels and C indicates a consonant. Arabic utterances can only start with a consonant [2]. Table 1 shows the Arabic alphabets along with number and types of syllables in every spoken alphabet.

Table 1: Arabic Alphabets

Alphabet	Arabic Writing	Syllables	No. of Syllables	Alphabet	Arabic Writing	Syllables	No. of Syllables
Alef	ألف	CV-CVC	2	Dhaad	ضاد	CVC	1
Hamzah	همزة	CVC-CVC	2	T_aa	طا	CV	1
Baa	با	CV	1	Dhaa	ظا	CV	1
Taa	تا	CV	1	Ain	عين	CV-CVC	2
Thaa	ثا	CV	1	Ghain	غين	CV-CVC	2
Jeem	جيم	CVC	1	Faa	فا	CV	1
H_aa	حا	CV	1	Qaaf	قا ف	CVC	1
Khaa	خا	CV	1	Kaaf	كا ف	CVC	1
Daal	دال	CVC	1	Laam	لام	CVC	1
Thaal	ذال	CVC	1	Meem	ميم	CVC	1
Raa	را	CV	1	Noon	نون	CVC	1
Zain	زين	CV-CVC	2	Haa	ها	CV	1
Seen	سين	CVC	1	Wawo	واو	CVC	1
Sheen	شين	CVC	1	Yaa	يا	CV	1
Saad	صاد	CVC	1				

1.2 Spoken alphabets Recognition

In general, spoken alphabets 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 Markov 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 alphabets recognition. In 1985, Hagos [9][4] 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 alphabets 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.

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.

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 alphabets 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 alphabets 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.

SAAVB contains three subsets called SAAVB26, SAAVB27, and SAAVB28 as following [15]. First, SAAVB26 contains read speech as spelling of a phonetically rich word (671 randomized and partially repeated items). Second, SAAVB27 contains read speech as spelling of a city name (Prompt 20 and 21). Last, SAAVB28 contains read speech as spelling of a personal name (260 randomized and repeated items). Each alphabet is supposed one of the Arabic alphabets which contains about 35 different alphabets, but we used in our final system testing 29 alphabets.

3 Results

We conducted different experiments and got four confusion matrices. The system was trained and tested by only SAAVB26, SAAVB27, SAAVB28 separately and, then, combined these three (SAAVB26, SAAVB27, SAAVB28) named as SAAVBaa. In all three experiments training and testing subsets are disjoint and the number of tokens (alphabets) in data subset for testing in SAAVB26, SAAVB27, and SAAVB28 were 923, 1,594 and 1,110 respectively and the number of tokens in data subset for training was about three times of their respective testing data subset.

As can be seen from Table 2, the correct rates of the system are 62.84%, 67%, and 61.26% for using SAAVB26, SAAVB27, and SAAVB28, respectively. The worst correct rate was encountered in the case of SAAVB28 while the best correct rate was encountered in the case of SAAVB27. We think this is correlated to the amount of the training data subset because SAAVB26 and SAAVB28 have low training data compared to SAAVB27. As can be noticed here the correct rate and size of SAAVB28 is in between if compared to that of SAAVB26 and SAAVB27. We want to emphasize that the size of training subset is three times the testing subsets in all cases.

By mixing training subsets of SAAVB26, SAAVB27, and SAAVB28 in one training subset and mixing all testing subset of these portions in one testing subset we get the set called SAAVBaa, where the confusion matrix for this set is shown in Table 3. Depending on testing this subset (i.e., SAAVBaa), the system must try to recognize 3,623 samples for all 29 alphabets. The overall system performance was 64.06%, which is reasonably high where our database is considered as noisy corpus.

The similarity between groups of Arabic alphabets is very high. For example we may have two alphabets with the only difference is the voicing in the carried phoneme such as the case of alphabets A4 and A9. We have many cases of these pair of phonemes that can cause misleading behavior of the system and we can notice such problems in the confusion matrices. The system failed in recognizing total of 1,342 alphabets (1,302 were substituted and 40 were deleted mistakenly by the system) out

of 3,623 recorded alphabets. Alphabets A7, A15, A23, A28, and A29 have gotten reasonably (above 85%) high recognition rate; on the other hand, the bad performance was encountered with alphabets A4, A10, A20, A21, A25, and A27 where the performance is less than 50%. Even though the database size is medium (only the 29 spoken Arabic alphabets) and with the existence of noise, the system showed a reasonable performance due to the variability in how to pronounce Arabic alphabets.

Table 2: Summary of accuracies for all subsets and alphabets

Arabic	Symbol	SAAVB26	SAAVB27	SAAVB28	SAAVBaa
أ	A1	25	53.99	34.03	60.96
هـ	A2	58.82	33.33	25	56.67
إ	A3	36.11	58.24	40	50.91
آ	A4	20	33.8	13.33	23.39
أ	A5	23.81	0	0	7.41
م	A6	50	73.81	44	66.28
ن	A7	70.59	58.82	84.62	86.67
و	A8	86.21	71.43	0	82.98
د	A9	76.19	91.38	90.91	80.84
ذ	A10	14.29	0	0	8
ر	A11	60.42	74.07	70.15	56.56
ز	A12	68.75	7.14	45.83	50
ع	A13	72.73	75.41	67.74	70.18
ث	A14	56.25	87.5	70	66.04
ل	A15	73.91	50	50	85.37
ط	A16	100	100	100	100
ظ	A17	57.14	52.94	36.36	64.58
ظ	A18	100	100	100	100
ع	A19	72	75	69.81	74.75
س	A20	78.95	30	44.44	47.37
ق	A21	62.79	72	66.67	39.05
ق	A22	89.47	85.71	80	81.16
ك	A23	95.83	83.02	86.67	90.22
ل	A24	93.18	85.51	80.28	83
م	A25	54.55	32.56	20.88	27.08
ن	A26	81.63	73.64	80.82	79.91
هـ	A27	25	51.52	34.43	26.03
و	A28	95.56	97.69	97.87	96.4
ي	A29	82.5	85.83	92.38	85.9
Overall (%)		62.84	67	61.26	64.06

Table 3: Confusion matrix of the system when trained and tested with SAAVBaa

Column	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	
A1	278	85	5	4	0	2	4	3	3	1	1	4	5	5	6	0	1	1	1	0	15	0	7	7	1	0	3	2	
A2	5	34	1	2	0	0	1	0	1	0	0	2	1	1	1	2	2	0	0	0	4	0	2	0	0	0	0	1	
A3	1	5	84	1	0	1	7	0	33	1	8	3	0	1	1	0	1	1	2	0	4	2	3	2	0	0	2	1	
A4	6	14	4	29	2	1	2	1	1	0	2	0	5	0	2	1	1	1	0	0	12	0	26	2	1	0	2	2	
A5	3	2	0	1	2	0	0	2	0	0	2	0	3	0	1	0	0	0	0	0	3	0	6	0	0	1	0	0	
A6	1	3	1	1	0	57	0	0	0	1	1	1	8	0	0	0	0	1	0	2	2	1	1	1	1	0	0	1	
A7	0	1	0	0	0	0	78	1	0	1	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
A8	0	0	1	0	0	0	39	0	1	1	0	0	0	1	0	0	0	0	0	0	3	0	0	0	0	0	1	0	
A9	1	5	2	4	0	0	2	135	1	0	2	0	0	0	0	0	1	1	1	1	6	0	1	2	0	0	0	0	
A10	1	3	2	1	0	1	0	0	5	2	0	3	0	1	0	0	1	0	0	0	2	0	0	1	0	0	0	1	
A11	6	13	3	6	2	0	1	2	3	2	125	1	5	1	7	3	13	4	0	0	7	4	2	2	0	0	3	3	
A12	2	2	3	0	0	3	1	0	3	0	1	27	1	1	0	0	1	0	1	0	3	2	0	1	0	0	1	0	
A13	3	4	0	2	0	7	0	3	0	0	0	0	80	4	1	0	1	0	1	0	1	3	1	1	0	0	0	0	
A14	1	0	0	0	0	3	0	1	0	0	0	1	5	35	1	0	0	0	0	2	0	0	3	0	1	0	0	0	
A15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35	0	1	1	0	0	1	0	1	0	1	0	0	0	
A16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A17	1	1	0	2	0	0	0	0	0	0	2	0	0	1	2	0	31	0	0	0	0	0	6	1	0	0	0	1	0
A18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A19	1	4	0	0	0	3	1	0	1	1	0	0	1	0	1	0	0	0	0	74	3	2	0	1	1	0	1	1	0
A20	0	0	2	0	0	2	0	1	0	0	4	1	2	0	1	0	1	0	3	18	1	0	0	0	0	0	0	1	0
A21	2	4	1	3	3	0	3	3	0	0	2	1	24	1	4	0	1	0	1	1	41	2	3	0	0	0	1	0	2
A22	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	3	1	1	0	3	56	1	0	0	0	0	0	0
A23	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	83	0	0	0	0	0	0	0
A24	1	3	2	0	0	1	4	0	6	3	1	8	0	0	0	0	2	0	1	0	4	0	0	210	0	2	0	1	
A25	4	11	3	6	2	60	3	3	11	2	1	2	15	2	2	0	1	0	1	3	21	3	0	7	65	2	5	0	
A26	0	8	0	0	0	0	1	6	0	0	1	0	2	0	0	0	1	1	9	0	2	8	0	183	3	0	0	0	
A27	4	14	4	6	0	0	17	8	2	5	5	3	3	1	3	0	3	8	1	1	10	4	0	3	0	0	38	2	
A28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	3	0	0	1	0	0	1	214	

Number of Alphabet tokens appeared in the test data subset from 0 up to 469 times. This is one of the drawbacks of the SAAVB26, SAAVB27, SAAVB28 subsets. Some of the alphabets were completely disappeared from the vocalized version due to problems of misleading effect in selecting and/or bad vocalizations given by speakers. A16 (ض) and A18 (ظ) are two of the alphabets that appeared with 0 times in all three subsets. Also the inconsistency of number of occurrence of alphabets in each subset of SAAVB26, SAAVB27, and SAAVB28 showed a fatal disadvantage of SAAVB which may caused a biased training among different system vocabulary. To give an example, alphabet A1 appeared about 1407 times in all three subsets while A10 and A18 appeared 24 and 0 times in all three subsets, respectively.

In our experiments, we can notice that some of the alphabet got a very bad accuracy in all subsets which mean that the problem of this is caused low training data regarding this alphabet and/or the high similarity of this alphabet to another alphabet(s). Examples of this situation are A4 and A27. Also it can be noticed that some of the alphabets got a very low accuracy in one subset but gained a high accuracy in case of the others. Example of this situation is A20.

4 Conclusion

To conclude, a spoken Arabic alphabets recognizer is designed to investigate the process of automatic alphabets 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. There are total of four experiment were conducted on Arabic alphabets. The SAAVB corpus (especially subsets dedicated for isolated Arabic alphabets, namely, SAAVB26, SAAVB27, and SAAVB28) supplied by KACST is used in this research. The first three experiments were conducted on each

of these three subsets and the fourth one was conducted on sum of all these subsets (i.e., by using all data of alphabet in one experiment). The system correct rate for the grand one is 64.06%.

5 Acknowledgment

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

References

- [1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers and http://en.wikipedia.org/wiki/Arab_world.
- [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 Science 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/htkbook.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.