Towards Using CMU Sphinx Tools for the Holy Quran Recitation Verification

Mohamed Yassine El Amrani[1,2,a], M. M. Hafizur Rahman[2,b],
Mohamed Ridza Wahiddin[2,c], and Asadullah Shah[2,d]

[1] Jubail University College, Saudi Arabia

[2] International Islamic University Malaysia, Malaysia

[a] amranim@ucj.edu.sa, [b] hafizur@iium.edu.my, [c] mridza@iium.edu.my, [d] asadullah@iium.edu.my

ABSTRACT

The use of the Automatic Speech Recognition (ASR) technology is being used in many different applications that help simplify the interaction with a wider range of devices. This paper investigates the use of a simplified set of phonemes in an ASR system applied to Holy Quran. The Carnegie Mellon University Sphinx 4 tools were used to train and evaluate a language model on Holy Quran recitations that are widely available online. The building of the language model was done using a simplified list of phonemes instead of the mainly used Romanized in order to simplify the process of training the acoustic model. In this paper, the experiments resulted in Word Error Rates (WER) as low as 1.5% even with a very small set of audio files used during the training phase.

***Index Terms*—Automatic speech recognition, Holy Quran recognition, Human voice.**

I. I 1 ntroduction

The Holy Quran is at the center of every Muslim in the world. Practicing 5 daily prayers is one of the five pillars of Islam and every prayer involves some Holy Quran recitation. Hence, every Muslim is involved in some Holy Quran memorization in order to recite some *Ayat* ("sentences" in the Holy Quran) during his/her daily prayers. Many Ayat encourage memorizing the Holy Quran and the mention of its easiness of memorization is repeated several times as in the *Ayah*:

{ وَلَقَدْ يَسَّرْنَا الْقُرْآنَ لِلذِّكْرِ فَهَلْ مِن مُّدَّكِرٍ }

"And We have certainly made the Quran easy for remembrance, so is there any who will remember?" (50:17)

As of any memorization process, a continuous review is required to remember as what was previously memorized. The best way to have one's recitation reviewed and verified is to involve another person who follows one's recitation and correct his mistakes. When no one is available to check someone's recitation, the verification cannot be done easily since the reciter has to continuously check after each group of *Ayat* if his recitation was correct. In this case, the reciter needs to look at the Holy Quran Book while trying to avoid reading the next *Ayat* in order not to benefit from a hint on the next Ayah ("sentence" in the Holy Quran) to recite.

The advances in Automatic Speech Recognition (ASR) should allow Quran reciter to not completely rely on another person to check the correctness of his recitation but also, the wide availability of devices equipped with microphones and powerful computing capabilities constitute a great opportunity for using ASR and help Quran reciters check their recitations.

Nowadays, many users use their mobile devices to read the Holy Quran. The purpose of this research is to help simplify the process of building a robust acoustic model for the Holy Quran recitation verification of both the Arabic and Tajweed rules.

The remainder of this article is organized as follows: A brief description of the Automatic Arabic Speech Recognition (AASR) applied to the Holy Quran is discussed in section 2. The experimental setup for using the Sphinx tools to build the acoustic model for the Holy Quran is introduced in section 3. The results of the experiments are discussed in the section 4. Finally, the conclusion and future work are discussed in section 5.

2 Automatic Arabic Speech Recognition (AASR)

There are many tools that can be used for the AASR which can be applied to the Holy Quran. One of the most actively researched tools is the Carnegie Melon University (CMU) Sphinx which is a statistical speaker-independent set of tools using the Hidden Markov Models (HMM). Many researchers have been using the Sphinx tools for Arabic speech recognition as well as for the Holy Quran as shown in (Abushariah, Ainon, Zainuddin, Elshafei, & Khalifa, 2010, 2012; Elshafei, Al-muhtaseb, & Al-ghamdi, 2008; Lamere et al., 2003; Satori, Harti, & Chenfour, 2007; Walker et al., 2004). The idea of applying the ASR techniques to the Holy Quran is not new and the Holy Quran is the subject of many research works. Researchers worked on different tasks like improving the correct pronunciation of letters as in (Arshad et al., 2013), the recognition of Tajweed rules found in (Ibrahim, Zulkifli, & Razak, 2011), the creation of a virtual learning systems as in (Mohamed, Hassanin, & Ben Othman, 2014) and (Yekache, Kouninef, Mekelleche, & Mohamed, 2013), and the detecting the mistakes of students' recitations like in (Ahsiah, Noor, & Idris, 2013).

An introduction to CMU Sphinx 4 for the Arabic Speech Recognition is provided in (Lamere et al., 2003; Satori et al., 2007; Walker et al., 2004). HMM statistical language model contains uni-grams, bi-grams, and tri-grams probabilities. This has the advantage of allowing speaker-independent models to be trained quickly with accurate recognition rates.

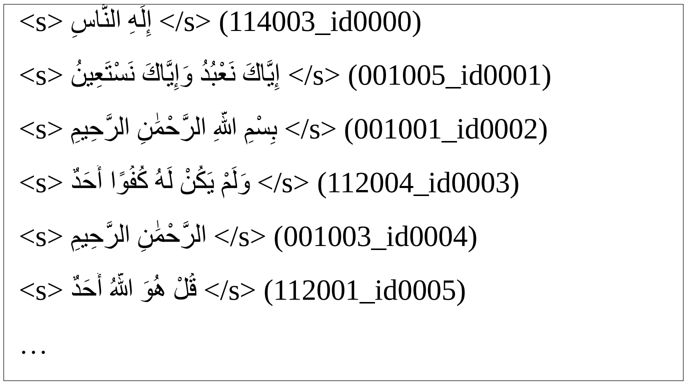
Many researchers in the field of AASR are using a Romanized set of phonemes in order to train their acoustic model. While this use obtained good results, the process of preparing and generating the phonemes set used in the phonetic transcription of the audio files used in the training of the acoustic model. In this paper, the focus will be to investigate the performance of the use of a simplified set of Arabic phonemes instead of the Romanized set of phonemes.

3 Acoustic Model for the Holy Quran

The CMU Sphinx 4 training algorithm requires several data in order to build an acoustic model to use for the speech recognition: a transcription file, a phonetic dictionary, a set of phonemes, and the audio files. The following is the description of the main data used to train the acoustic model for 4 Holy Quran *Surat* (chapter): 001, 112, 113, and 114. The choice of those chapters was motivated by the fact that they are very short in term of the length of the *Ayat* and many people are memorizing them.

3.1 Transcription file

The transcription file links the Arabic *Ayat* to their respective audio files. It consists of the transcription of the content of each audio file of the Arabic *Ayah* between the delimiters “<s>” and “</s>” along with the unique identifier of the audio file as shown in figure 1. In some audio files where there is an adult reciter and a child repeating after him the recited *Ayah*, an automated processing was added in order to tag with “<sil>” the silence between the two recitations.



```
<s> إِلَٰهِ النَّاسِ </s> (114003_id0000)
<s> إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ </s> (001005_id0001)
<s> بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ </s> (001001_id0002)
<s> وَلَمْ يَكُنْ لَهُ كُفُوًا أَحَدٌ </s> (112004_id0003)
<s> الرَّحْمَٰنِ الرَّحِيمِ </s> (001003_id0004)
<s> قُلْ هُوَ اللَّهُ أَحَدٌ </s> (112001_id0005)
…
```

Figure 1. Sample of a transcription file for the chapters 001, 112, 113, and 114 of the Holy Quran

The audio files are renamed and indexed using the following template: “*CCCAAA_idNNNN*” where *CCC* is the chapter number, *AAA* the *Ayah* number, and *NNNN* the sequential number; as a unique identifier; for each file since there is more than one recitation of the same *Ayah* of the same chapter.

This is automatically generated from a MySQL database holding the Holy Quran using the audio recitation files widely publicly available in (“Download Quran Text - Tanzil,” n.d.) and (“Verse By Verse Quran Files,” n.d.).

3.2 Phonetic dictionary

The phonetic dictionary consists of the list of phonemes that constitutes the word as is it spoken in the associated audio file. This is also automatically generated using a simplified representation of the phonemes. The pronunciation of the word varies depending on its position in the *Ayah*.

For example: (الرَّحْمَٰنِ) (The most merciful) appears in two different positions on different *Ayat*, hence it has two sets of phonemes, since its pronunciation differs when it is in the middle or in the beginning of the *Ayah*.

The result of the automatic generation of the phonetic dictionary is shown in table 1.

Table 1
Example of phonetic transcription for the chapters 001, 112, 113, and 114 of the Holy Quran

بِسْمِ	بِ سْ مِ
اللَّهِ	لَّ هِ
الرَّحْمَٰنِ	رَّ حْ مَ ا نِ
الرَّحْمَٰنِ(2)	ءَ رَّ حْ مَ ا نِ
…	
الْجِنَّةِ	لْ جِ نَّ تِ
وَالنَّاسِ	وَ نَّ ا سْ

3.3 Arabic phonemes

The idea of using a list of phonemes using the Arabic letters was mentioned by researchers in previous articles like in (Brierley, Sawalha, Heselwood, & Atwell, 2014) and (Yekache, Mekelleche, & Kouninef, 2012). However, either no performance of the acoustic model was supplied or the Arabic phonemes were only mapped to the Romanized phonemes.

Table 2
Phonemes generated for the chapters 001, 112, 113, and 114 of the Holy Quran

1	ءَ	17	دِ	33	صِ	49	قَ	65	نَ
2	ءِ	18	دِّ	34	صَّ	50	قُ	66	نِ
3	ا	19	دْ	35	صِّ	51	قِ	67	نَّ
4	بُ	20	ذَ	36	ضُ	52	قْ	68	نْ
5	بِ	21	ذُ	37	ضَّ	53	كَ	69	هُ
6	بِّ	22	ذِ	38	طَ	54	كُ	70	هِ
7	بْ	23	رَ	39	عَ	55	كِ	71	هْ
8	تَ	24	رِ	40	عُ	56	لَ	72	و
9	تِ	25	رَّ	41	عِ	57	لِ	73	وَ
10	ثَ	26	رِّ	42	عْ	58	لَّ	74	وِ
11	جِ	27	سَ	43	غَ	59	لِّ	75	وْ
12	حَ	28	سُ	44	غْ	60	لْ	76	ي
13	حِ	29	سِ	45	فَ	61	مَ	77	يَ
14	حْ	30	سْ	46	فُ	62	مُ	78	يُ
15	خَ	31	شَ	47	فِ	63	مِ	79	يَّ
16	دُ	32	صُ	48	فَّ	64	مْ	80	يْ

The set of phonemes used in this investigation (see table 2) for the phonetic dictionary consists of an Arabic letter along with its diacritics. In addition, “ ا ”, and “ ي ”, and “ و ” are the only 3 Arabic letters without diacritics and are used for the elongations of a letter with *Fathah* “◌َ”, *Kasrah* “◌ِ”, and *Dammah* “◌ُ” respectively. The emphasis symbol (*Shaddah:* “◌ّ”) can be also found with a diacritic symbol in order to represent the emphasis on the pronunciation of a letter. Finally, the *Hamzah* “ ء ” has been used to represent all its different forms (ا أ إ ئ ؤ ء) depending on its diacritic or position in the sentence.

This simplification of the phonemes allows verifying its correctness without any experience which leads to a faster acoustic model building. This simplification allows capturing all the different Arabic and Tajweed rules.

3.4 Training data

The audio files used for the building of the acoustic model consist of recordings of recitations of every *Ayah* of the selected 4 chapters (001, 112, 113, and 114) from (“Verse By Verse Quran Files,” n.d.). These recordings of speech utterances were converted to 16 kHz, 16 bit, mono files, in MS WAV format as this is required by the Sphinx trainer (version 1.08). Some manual processing had to be done by trimming the beginning and the end of the audio files to match the *Ayah* since it might include some utterance not present in the transcription from the Holy Quran or because of the existence of long silences that needed to be removed. Also, transcriptions of some relatively long *Ayat* needed to be manually updated in order to add a silence tag when the reciter pauses his recitation in the middle of the *Ayah*.

The estimated total length of the training audio files is about 40 minutes. Even if this is a small amount of audio data and all of the audio files were used for both the training and the testing phases, the results were very promising. Those audio files are all from male speakers consisting of 22 adults and 1 child.

4 Results and discussion

Using the previously described data for the 65 words and 80 phonemes for the selected 4 chapters of the Holy Quran, many different training configurations were tried while varying the number of tied states (senones) and the dimension of the Gaussian Mixtures since they have a direct impact on the performance of the speech recognition. While the number of senones adjusts the context-dependency of each audio signal associated to a phoneme, the dimension of the Gaussians define the number of Gaussians used in the Gaussian Mixture approximating a subset of a sound wave.

The training scenarios were generated by varying the number of senones from 100 to 2000 with an increment of 50. For each number of senones used, the dimension of the Gaussian was set to a value from 2 to 64 with an increment of 2. Finally, the training was repeated twice for each combination of a number of senones and a dimension. The average of the performance of the training is calculated using the Word Error Rate (WER) provided by the Sphinx testing tool. The repetition of the training was made since in some occurrences, the same settings led to slightly different WERs. The results obtained (figure 2) shows the obtained average WER for each training scenario. It confirms that it is possible to achieve very good word recognition accuracy with a simplified set of phonemes with many different settings reaching a WER lower than 2%. It shows that setting the dimension to 32 is the optimal value for many different numbers of senones allowing reaching the minimum WER of 1.5% with 1850 as the number of senones.

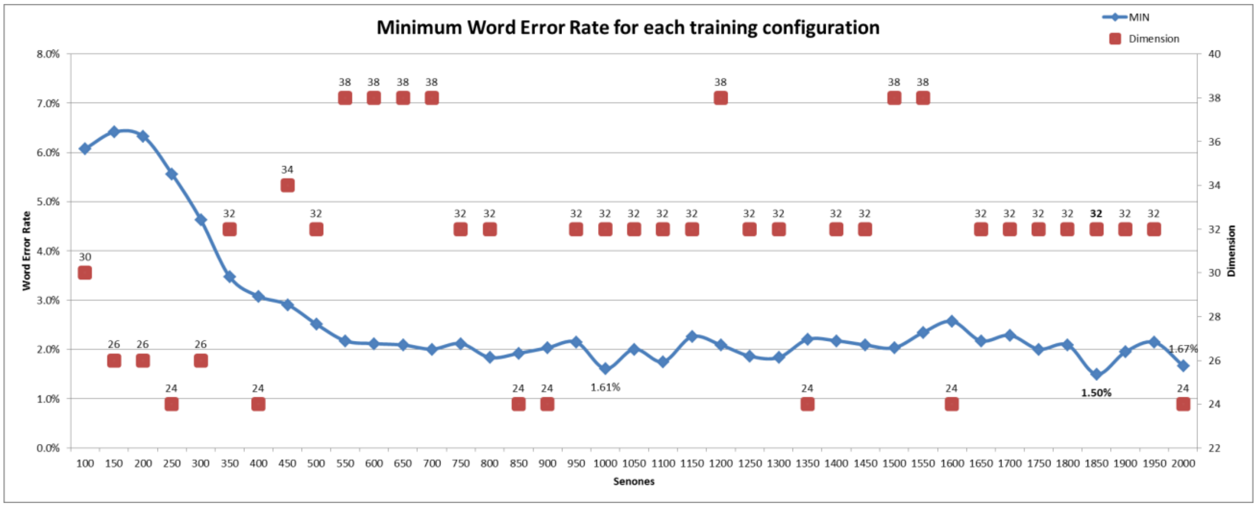


Figure 2. The dimension of the Gaussians that led to the minimum value of the average WER for each setting of the number of senones

5 Conclusion

In this paper, the use of a simplified set of Arabic phonemes for the Arabic Automatic Speech Recognition using the CMU Sphinx tools for the Holy Quran was investigated. The results obtained show that it is possible to obtain very interesting recognition accuracy with this simplified list of phonemes instead of using the Romanization method which is more elaborate to generate. The WER obtained for a number of senones of 1850 and the dimension of Gaussians of 32 was 1.5%. Future work will extend the generation of the phonetic transcription to more Holy Quran chapters to ultimately include all chapters along with adding more recitations from different renowned scholars in order to increase the training data.

References

Abushariah, M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2010). Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools. In *International Conference on Computer and Communication Engineering* (pp. 1–6). IEEE. doi:10.1109/ICCCE.2010.5556829

Abushariah, M., Ainon, R., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2012). Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus. *The International Arab Journal of Information Technology*, *9*(1), 84–93.

Ahsiah, I., Noor, N. M., & Idris, M. Y. I. (2013). Tajweed checking system to support recitation. In *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 189–193). IEEE. doi:10.1109/ICACSIS.2013.6761574

Arshad, N. W., Sukri, S. M., Muhammad, L. N., Ahmad, H., Rosyati Hamid, Naim, F., & Naharuddin, N. Z. A. (2013). Makhraj Recognition for Al-Quran Recitation using MFCC. *International Journal of Intelligent Information Processing*, *4*(2), 45–53. doi:10.4156/ijiip.vol4.issue2.5

Brierley, C., Sawalha, M., Heselwood, B., & Atwell, E. S. (2014). A Verified Arabic-IPA Mapping for Arabic Transcription Technology, Informed by Quranic Recitation, Traditional Arabic Linguistics, and Modern Phonetics. *Journal of Semitic Studies*.

Download Quran Text - Tanzil. (n.d.). Retrieved July 21, 2015, from http://tanzil.net/download/

Elshafei, M., Al-muhtaseb, H., & Al-ghamdi, M. (2008). Speaker-Independent Natural Arabic Speech Recognition System. In *The International Conference on Intelligent Systems*.

Ibrahim, N. J., Zulkifli, M. Y., & Razak, Z. (2011). Improve Design for Automated Tajweed Checking Rules Engine of Quranic Verse Recitation: A Review. *QURANICA-International Journal of Quranic Research*, *1*(1), 39–50.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., … Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong* (Vol. 1, pp. 2–5).

Mohamed, S. A. E., Hassanin, A. S., & Ben Othman, M. T. (2014). Virtual Learning System (Miqra ' ah) for Quran Recitations for Sighted and Blind Students. *Journal of Software Engineering and Applications*, (April), 195–205.

Satori, H., Harti, M., & Chenfour, N. (2007). Introduction to Arabic Speech Recognition Using CMUSphinx System. In *Proceedings of Information and Communication Technologies International Symposium*. Retrieved from http://arxiv.org/abs/0704.2083

Verse By Verse Quran Files. (n.d.). Retrieved July 22, 2015, from http://www.everyayah.com/data/status.php

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., … Woelfel, J. (2004). Sphinx-4 : A Flexible Open Source Framework for Speech Recognition. *SMLI*, (TR-2004-139), 1–9. doi:10.1.1.91.4704

Yekache, Y., Kouninef, B., Mekelleche, Y., & Mohamed, S. (2013). Building Quranic reader voice interface using sphinx toolkit. *Journal of American Science*, *9*(11), 473–479.

Yekache, Y., Mekelleche, Y., & Kouninef, B. (2012). Towards Quranic reader controlled by speech. *International Journal of Advanced Computer Science and Applications*, *2*(11), 134–137. Other Computer Science. Retrieved from http://arxiv.org/abs/1204.1566