International Journal on Islamic Applications in Computer Science And Technology, Vol. 3, Issue 4, December 2015, 17-28

Automatic Rule Based Phonetic Transcription and Syllabification for Quranic Text

Sameh A. Bellegdi[1], Husni A. Al-Muhtaseb[2]
[1]Academic Leadership Center, Ministry of Education, Saudi Arabia
[2]Information and Computer Science Department, KFUPM, Dhahran 31261, Saudi Arabia
[1]bellegdi@kfupm.edu.sa, [2]muhtaseb@kfupm.edu.sa

ABSTRACT

Speech processing has been the subject of an extensive number of research studies. Speech synthesis is the process of transferring text to speech. Phonetic transcription represents an essential part of any text-to-speech system. This paper proposes a transcription technique dedicated for the Quranic text. Transcribing Quranic text is a challenging problem as some letters have different phonemes for the same letter, depending on its neighbors. Different rules are proposed to handle the problem of Quranic text transcription depending on the art of Intonation (Tajweed). In addition, a rule based syllabification technique is presented. This research will have a good impact in the service of Holy Quran and its science. This research work is important to implement Quran recitation synthesis prototype as it addresses Quranic text transcription and syllabification. Quran recitation synthesis has main motivation of reducing space of Quranic sound files.

***Keywords*: Holy Quran, Quranic text transcription, Quranic text syllabification, text phonetization, grapheme-to-phoneme, Arabic TTS.**

1. INTRODUCTION

Text transcription – phonetization or grapheme to phoneme conversion – is a process of converting written text to phonetic transcription. Text transcription is an important step for many speech processing applications including text–to–speech synthesis and speech recognition. Text syllabification is a process of breaking text into syllables by correctly detecting syllable boundaries.

Transcribing Arabic text requires that the given text is fully vocalized (proper diacritics are placed on each letter of the text). Comparing with normal Arabic text, Quranic text transcription has more challenges as there are some letters which have different phonemes for the same letter depending on the preceding or succeeding letter(s). In addition, some text letters are not pronounced.

Different researchers have addressed the problem of Arabic text transcription and syllabification (Imedjdouben & Houacine, 2013), (Ramsay, Alsharhan, & Ahmed, 2014), (Imedjdouben & Houacine, 2014), (Soori, Platos, Snasel, & Abdulla, 2011). However; to the best of our knowledge, there is no work dedicated to address Quranic text transcription and syllabification. This paper addresses transcription and syllabification problems of the Quranic text by proposing a rule-based technique.

Quran recitation synthesis has main motivation of reducing space of Quranic sound files. Synthesizing Quran recitation requires many modules; viz. text analysis, phonetic analysis, prosodic analysis, and speech synthesis. This paper addresses text analysis module.

The rest of this paper is organized as follows. Background on the Quranic text and some of the used terminologies are presented in section 2. Section **Error! Reference source not found.** presents related work. Section 4 describes the proposed transcription and syllabification technique. Conclusion and future work are presented in Section 5.

2. BACKGROUND

2.1 Quranic Text Characteristics

Quranic text is a form of classical Arabic text. Quran is used as a reference for Arabic language (Harrag & Mohamadi, 2010). Holy Quran is a common element among Muslims all around the globe. The text of the Holy Quran is fully vocalized. Moreover, Quranic text has different symbols and special characters to guide readers for correct recitation. The accuracy of Quranic text is critical as it is the core of correct recitation.

2.2 Terminologies

The terms used in this paper are as follows:

- Vowels: There are two types of vowels: short vowels: *Fat-ha*, *Kasra*, and *Dhamma* that were written as "ـَ ـِ ـُ", respectively; and long vowels: (ا,ـُو,ـِي). It is worth mentioning that the letters (و، ي) may also be consonants.
- *Sukoon*: absence of a vowel.
- *Shaddah*: doubled letter.
- *Noon Saakinah*: is a *Noon* (ن) letter with no diacritic mark or with a *Sukoon*.
- *Meem Saakinah*: is a *Meem* (م) letter with no diacritic mark or with a *Sukoon*.
- Nunation (*Tanween*): it is a *Noon Saakinah* appears at the end of some words and is written as double *Dhamma*, double *Fat-ha*, or double *Kasra*.
- Nasal tone (*Ghunnah*): it is a soft sound comes from the nose when pronouncing *Noon* and *Meem* (ن، م).
- *Hamzat Wasl*: it is a special form of the letter *Hamza* (ء) that appears at the beginning of some words. It is only pronounced when starting recitation by that word. Otherwise, it is not pronounced.

3. RELATED WORK

Several researchers have addressed text transcription problem (Pitakpawatkul, Suchato, Punyabukkana, & Wutiwiwatchai, 2013), (Schlippe, Ochs, & Schultz, 2012). Comparatively, fewer ones have addressed Arabic text transcription problem (Imedjdouben & Houacine, 2013), (Ramsay, Alsharhan, & Ahmed, 2014), (Imedjdouben & Houacine, 2014). In addition, no extensive work has been dedicated to address Quranic text phonetic transcription problem, to the best of our knowledge.

With respect to Arabic text transcription, Imedjdouben and Houacine presented a rule based text transcription technique for Arabic text (Imedjdouben & Houacine, 2013),

(Imedjdouben & Houacine, 2014). The authors defined a set of special words and abbreviations with their transcription. They used SAMPA (Speech Assessment Methods Phonetic Alphabet) notations (Raškinis, Raškinis, & Kazlauskienė, 2003), (Wells, 1997) in their transcription which is difficult to debug and understand by human being. The authors evaluated their techniques using Arabic corpus of sentences (Alghamdi, Alhamid, & Aldasuqi, 2003).

Ramsay et al. presented a two phase phonetic transcription technique using a set of rules (Ramsay, Alsharhan, & Ahmed, 2014). The phases include converting grapheme to phoneme and then phoneme-to-allophone. The authors validated their work in two stages. In the first stage, they applied their technique on the PENN Arabic Treebank Text corpus (Maamouri, Bies, Buckwalter, & Mekki, 2004) and obtained the transcription of a set of words. Then, for any chosen word, they found its occurrences in different contexts to ensure the transcription is the same. In the second stage, they chose a set of words and obtained their transcription. An annotator person was asked to markup recordings of these words by a set of native speakers. Then, they compared the manual transcription and the generated one.

When syllabification is considered, different researchers proposed syllabification algorithms for different languages (Hernández-Figueroa, Carreras-Riudavets, & Rodríguez-Rodríguez, 2013), (Eddington, Treiman, & Elzinga, 2013), (Ibrahim M. A., 2013). Limited research papers addressed the Arabic syllabification problem (Soori, Platos, Snasel, & Abdulla, 2011), (Elshafei, Al-Muhtaseb, & Al-Ghamdi, 2002). In addition, no detailed work has been published to Quranic text, as far as we know. Soori et al. presented an algorithm for syllabification of Arabic and its usage in text compression (Soori, Platos, Snasel, & Abdulla, 2011). The proposed algorithm seems to be inaccurate. Elshafei et al. defined different types of speech segments, viz. consonants–half vowels, half vowel–consonants, half vowels, middle portion of vowels, and suffix consonants (Elshafei, Al-Muhtaseb, & Al-Ghamdi, 2002). They defined different syllable patterns in the Arabic language: cV, cW, cVc, cWc, and cVcc, where c represents a consonant, V represents a vowel and W represents a long vowel.

Different research papers addressed the syllabification problem in different Arabic dialects (Ibrahim M. A., 2013), (Alhuwaykim, 2013), (Azmi & Tolba., 2008). Ibrahim presented an investigation about syllables and syllable patterns in English and Fawi dialect – related to Faw town in Iraq (Ibrahim M. A., 2013). He described differences and similarities between the studied languages in terms of syllable patterns and types. Alhuwaykim analyzed the syllabification of single intervocalic consonants in the Arabic dialect of Sakaka city – in Saudi Arabia (Alhuwaykim, 2013). He investigated how intervocalic consonants with different sonority profiles were treated. Azmi and Tolba proposed a syllable-based automatic speech recognition (ASR) systems of Arabic to be used in noisy environment (Azmi & Tolba., 2008). The authors reported that the recognition rate of ASR system using syllables outperformed ASR with monophones and triphones.

Many researchers proposed different techniques and tools for the Holy Quran recitation (Ibrahim, Idris, Razak, & Rahman, 2013) (Abdou, et al., 2006) (Abdo, Kandil, El-Bialy, & Fawzy, 2010) (Ahmed, 2004). Abdou et al. developed HAFSS© application as a tool for teaching of the correct recitation of the holy Quran (Abdou, et al., 2006). The author built an automatic generation of pronunciation hypotheses as part of the proposed system. They proposed a phoneme duration classification algorithm to detect recitation mistakes. In another work, Abdo et al. introduced a system for detecting some common pronunciation errors in

Quran recitation (Abdo, Kandil, El-Bialy, & Fawzy, 2010). The authors introduced a semi-automatic segmentation approach based on delta function of best Mel Frequency Cepstral Coefficients (MFCC). Similarly, Ibrahim et al. developed a Quranic verse recitation recognition system with tajweed checking rules function (Ibrahim, Idris, Razak, & Rahman, 2013). Their system allows a reciter to recite a part of the Quran and then it revises and corrects the recitation. Elhadj et al. presented an independent recognizer for allophonic sounds of the classical Arabic based on Quranic recitation (Elhadj, Alghamdi, & Alkanhal, 2013). The used speech sounds were extracted from recitations of a part of the Holy Quran of ten reciters. Speech signals are segmented and annotated manually into three levels, viz. words, phonemes, allophones. HMM with 3-emitting states was used. Every state has a continuous distribution with 16 Gaussian mixture distributions.

Ahmed studied the rules of Quran recitation as true acoustic phenomenon (Ahmed, 2004). He analyzed the voice of Sheikh AlHosary during reciting the Quran. The author tried to break the barrier between the theoretical and lingual side, and the practical side by analyzing Quran voices based on both; the modern scientific research and the rules of Tajeweed. The author described the levels of linguistic analysis and how Arabic scholars understand these levels. In addition, he studied the modern acoustic studies. Fundamental terms related to speech were described including phonetic, speech intensity, wave form, fundamental frequency, spectrogram, and transcription. The effect of Nasalization has been studied with Noon (ن) and Meem (م) with respect to letters' description and origination (Ahmed, 2004). The author studied the rules of these two letters with 215 verses from the Holy Quran from different Chapters. The total number of phonemes is 10527. Statistics of the rules of the two letters have been reported, including frequency, time duration, and acoustic properties. The author discussed the rules of Madd, which is stretching the sound of a vowel. He measured the time duration of Madd types using HTK tool and presented a comparative result.

4. TRANSCRIPTION AND SYLLABIFICATION

In this module, we preprocess the text to remove or replace some characters to streamline the transcription and syllabification processes. A set of preprocessing tasks are designed for this purpose. Table 1 shows the list of the characters and their corresponding Unicode values that are involved in the preprocessing tasks. The preprocessing module tasks are:

- Standardizing *Hamza* forms: replace all *Hamza* forms "أ, إ, ؤ, ئ" to a standard one "ء". This task is helpful for the process of transcription and it eliminates ambiguity as the phonetic representation for all forms is the same.
- Removing the doubled letter mark (*Shaddah*) from the first letter if any. This is because in Arabic it is impossible to start with a letter with *Shaddah*. The *Shaddah* is because of the preceding verse.
- Checking the last letter of a given verse and its diacritic mark. If the diacritic mark is *Tanween Fat-ha* "ً" and the vocalized letter is letter *Taa Marbota* "ة", replace the letter and the diacritic mark by letter *Haa* "ه" and *Sukoon* "ْ". However, if diacritic mark is *Tanween Fat-ha* and the vocalized letter is not *Taa Marbota* "ة", replace the diacritic mark with *Fat-ha* "َ" and *Alif*. Otherwise, replace the diacritic mark with *Sukoon* "ْ" if the last letter is not a long vowel. This process is designed such that the reciting system will stop at the end of each *Ayah*. This task is important as in Arabic the last letter should have *Sukoon*. In addition, *Tanween Fat-ha* is replaced by *Fat-ha* "َ" and *Alif* if the last letter is not *Taa Marbota*.
- Removing all pause marks because the assumption is that there is no stop except between different verses.

- Replacing end of *Aya* mark by newline code to differentiate between different verses.
- Removing *Madd* sign "ۤ" at the end of the given verse. This sign is related to the succeeded verse and since there is a stop between any two verses it will not be pronounced.

Removing small letters Waw and Yaa with Unicode values of (06E5, 06E6, and 06E7) at the end of a given verse. They are removed because their existence is depending on the succeeded verse. If they are in the middle, replace them with normal corresponding letters Waw and Yaa (و, ي). This is helpful for the process of transcription as it eliminates ambiguity between small letters and normal ones as the phonetic representation for them is same.

- Removing *Alif* of *Tanween Fat-ha* as it is not involved in the recitation.
- Replacing *Alif Maksoura* "ى" succeeded by a diacritic mark by the normal *Yaa* "ي". This is to eliminate ambiguity between letters *Alif* and *Yaa*. The letter *Alif* is not succeeded by any diacritic mark.
- Replace *Alif Maksoura* "ى" preceded by the diacritic mark *Kasra* by the normal *Yaa* "ي". Similar to the last task, this task eliminates ambiguity. It is impossible to have a latter *Alif* preceded by *Kasra*.
- Removing Small *Meem* "ۥ" with Unicode value of (06E2) if it comes at the end of a given verse. This character represents a rule for *Noon Sakinah* and *Tanween* (will be discussed later). In case of stopping at the end of a given verse, there will be no need for this character.
- Removing any *Alif* "ا" in the middle given that it is followed by the silent character "ٰ" with Unicode of (06E0). The silent character means that the preceded letter is not pronounced.
- Removing small *Seen* low "ۜس" with Unicode value of (06E3) .This character appears below letter *Sad* to indicate that it is possible to pronounce *Seen* instead of *Sad*. According to *Hafss* way, pronouncing *Sad* in this case is preferred.
- Replacing small *Noon* "ۢ" with Unicode value of (06E8) by the normal one. This is to eliminate ambiguity between small and normal ones as they are treated in the same way.
- Removing *Ishmam* sign "ۭ" with Unicode (06EB). This sign is removed because it is not involved in the pronunciation.

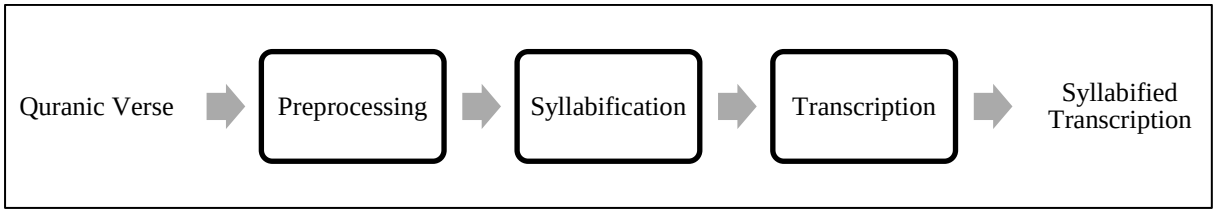

Figure 1. Proposed Prototype Architecture

Table 1: Unicode Values for Characters Involved in the Preprocessing

Code	Char	Name according to Unicode Table	Code	Char	Name according to Unicode Table
0621	ء	Arabic letter *Hamza*	0652	◌ْ	Arabic *Sukun*
0623	أ	Arabic letter *Alef* with *Hamza* above	0653	◌ٓ	Arabic *Maddah* above
0624	ؤ	Arabic letter *Waw* with *Hamza* above	06D6	◌ۖ	Arabic small high ligature sad with lam with *Alef Maksura*
0625	إ	Arabic letter *Alef* with *Hamza* below	06D7	◌ۗ	Arabic small high ligature *Qaf* with lam with *Alef Maksura*
0626	ئ	Arabic letter *Yeh* with *Hamza* above	06D8	◌ۘ	Arabic small high *Meem* initial form
0627	ا	Arabic letter *Alef*	06D9	◌ۙ	Arabic small high *Lam Alef*
0629	ة	Arabic letter *Teh Marbuta*	06DA	◌ۚ	Arabic small high *Jeem*
0646	ن	Arabic letter *Noon*	06DB	◌ۛ	Arabic small high three dots
0647	ه	Arabic letter *Heh*	06DE	۞	Arabic start of *Rub El Hizb*
0648	و	Arabic letter *Waw*	06E0	◌ۜ	Arabic small high upright rectangular zero
0649	ى	Arabic letter *Alef Maksura*	06E2	◌ۢ	Arabic small high *Meem* isolated form
064A	ي	Arabic letter *Yeh*	06E3	◌ۣ	Arabic small low *Seen*
064B	◌ً	Arabic *Fathatan*	06E4	◌ۤ	Arabic small high *Madda*
064C	◌ٌ	Arabic *Dammatan*	06E5	◌ۥ	Arabic small *Waw*
064D	◌ٍ	Arabic *Kasratan*	06E6	◌ۦ	Arabic small *Yeh*
064E	◌َ	Arabic *Fatha*	06E7	◌ۧ	Arabic small high *Yeh*
064F	◌ُ	Arabic *Damma*	06E8	◌ۨ	Arabic small high *Noon*
0650	◌ِ	Arabic *Kasra*	06EB	◌ۭ	Arabic empty centre high stop
0651	◌ّ	Arabic *Shadda*			

4.1 Syllabification

Choosing the right unit type and length for syllabification is an essential step in concatenative text-to-speech synthesis (Elshafei, Al-Muhtaseb, & Al-Ghamdi, 2002). There are different units that can be used to segment the Arabic text. Some of which are word-based, syllable-based, and diphone-based segmentations. In this work, we use an extended syllable-based segmentation for unit preparation.

Given ‘c’, ‘V’, and ‘W’ as consonant, vowel, and long vowel letters, respectively; Arabic syllable patterns can have these forms: cV, cW, cVc, cWc, cVcc, and cWcc (Elshafei, Al-Muhtaseb, & Al-Ghamdi, 2002), (Kiparsky, 2003).

Elshafei et al. (Elshafei, Al-Muhtaseb, & Al-Ghamdi, 2002) defined syllable components as *c, c’, x, x’, y, y’, z, z’, v, v’, w, w’*, where c: consonant, and c’: intervocal consonant. x: initial half vowel (HV), x’: initial accentuated HV, y: final HV, and y’: final accentuated HV. z: sustained (not initial) portion of a normal vowel, z’: sustained portion of an accentuated vowel, v: normal short vowel, v’: accentuated short vowel, w: a long vowel, and w’: accentuated long vowel. Different forms of every pattern with examples are given below

- cV: cx-y (“كـ” as in “كسب”), cx’-y’ (“غُـ” as in “غُصن”), cv (“أَ” as in “بدأ”), or cv’ (“قَ” as in “ضاقَ”) , cx-y’ (“كـ” as in “كَظم”), cx’-y (“صَـ” as in “صَعد”)
- cW: cx-z-y (“كا” as in “كان”), cx’-z’-y’ (“ضا” as in “ضاق”), cw (“عا” as in “ضاعا”), or cw’ (“غى” as in “أصغى”), cx’-z’-y (“ضا” as in “ضاع”), cx-z-y’ (“كا” as in “كاظم”)
- cVc: cx-yc (“تُبْ” as in “كُتُبْ”), cx’-y’c (“خَصْـ” as in “خَصْمٌ”), cx-y’c (“شَطْ” as in “شَطْرٌ”), or cx’-yc (“ظُلْـ” as in “ظُلْمٌ”)
- cWc: cx-z-yc (بَـابْ), cx'-z'-y’c (“صَاخَّـ” as in “صَاخَّة”), cx-z'-y'c (بَاقْ), or cx'-z'-yc (قَامْ)
- cVcc: cx-yc’-c (سَمَّ), cx'-y'c’-c (قُطْرْ), cx-y’c’-c (كَظْمْ), cx'-yc’-c (ظَنّ), all the given examples are in the case of stopping on the given word.
- cWcc: cWc’-c (جانّ) in the case of stopping on the given word.

Notice that we are extending Elshafei et al. (Elshafei, Al-Muhtaseb, & Al-Ghamdi, 2002) with these forms: cx-y’, cx’-y, cx’-z’-y, cx-z-y’ and cWc’-c.

Four conditions should be satisfied to determine the boundaries of a syllable:

1. The current character is a consonant letter and is not the last letter in the given verse text.
2. The next character is not *Sukoon* nor *Shaddah*.
3. If the current letter is followed by *Sukoon* or *Shaddah*, the next letter is not followed by *Sukoon* or *Shaddah*.
4. The next letter is not a vowel.

4.2 Transcription

The science of reading and the art of Intonation (*Tajweed*) are major factors in correct recitation. Taking these factors into account, we propose a set of rules for transcription. The proposed rules in this research work are based on *Hafss* way of reading (Alshatebi, 2014). These rules depend on the used *Othmani* Quranic text which has special characters for certain rules. More information about the art of *Tajweed* can be found in (Qamhawi, 1956).

Error! Reference source not found. shows an example of a transcription rule: the pseudo code of *Qalqala* rule.

A set of special words along with their syllabified transcription is defined. Table 2 shows the list of defined exception words. Majestic word of "الله" The God (*Allah*), is treated as a special word with two situations. If preceded by *Kasra* diacritic mark, it is softened. Otherwise, it is magnified. These situations are represented by special identifying marks. For more clarity, see example 2 in Table 3. Some of the special words have an extended *Madd*. There are two types of *Madd* depending on the diacritic mark follows the long vowel, that are *Sukoon* or *Shaddah*. If the diacritic mark is *Shaddah*, the case is represented by “~5”. Otherwise, it is represented by “~6”. See examples 4 and 5 in Table 3.

The general idea of the transcription is that any consonant letter is followed by a diacritic mark or *Sukoon*. However, for the given letters in the coming rules, special identifying marks are used instead of *Sukoon* to represent the type of the satisfied rule. In addition, if the letter is not followed by any diacritic mark nor *Sukoon*, this means that the rule is partial Merging (*Idgham Naqis*) except in the case of long vowels. As an example, the word “بَسَطتَ” has letter “ط” that is not vocalized nor followed by *Sukoon*. As a result, any letter in the transcribed text has its special phoneme depending on its diacritic mark or the given identifying marks.

In this paper, different rules are proposed based on the art of *Tajweed* “Intontation” to

perform the transcription. Note that the rules for "Stopping *Sukoon* extension" is not included. The proposed rules are as follows:

4.2.1 Noon Sakinah and Tanween:

This rule is concerned about *Noon Sakinah* {نْ} and nunation (*Tanween*). Let us denote *Noon Sakinah* letter or *Tanween* by N. In addition to the normal case, there are three situations for this rule; viz. Merging (*Idgham*), Hiding (*Ikhfaa*), and Replacing (*Iqlab*) depending on the letter that succeeds N. If the letter is one of {ي، و}, the rule is Merging. If this letter is one of these letters {ث، ج، د، ذ، ز، س، ش، ص، ض، ط، ظ، ف، ق، ك، ت}, then the rule is Hiding. If it is {ب}, N is converted to *Meem Sakinah* and follows the next rule *Meem Sakinah*. In all three cases, a special mark is used with N to denote *Ghunnah*. Finally, if the letter is one of the letters {ل، ر}, then N is removed as the rule is full merging. If N is succeeded by letter {ن}, the rule of *Noon and Meem Mushadda* is applied (will be described later).

```
If the current letter is Qalqala letter
    If the current letter is the last one in the verse (Aya)
       If followed by Shaddah
          Transcription  = Transcription + current letter + Qalqala Kubra Mark
       Else if followed by Sokoon
          Transcription  = Transcription + current letter + Qalqala Wusta Mark
       End if
   Else
         Transcription  = Transcription + current letter + Qalqala Sughra Mark
    End if
End if
```

Figure 2. An example of transcription rules: Qalqala rules

Table 2: List of Special Words

الٓمٓصٓ	الٓمٓ	بِسۡمِ	ٱللَّهِ
كٓهيعٓصٓ	الٓمٓر	مَجۡرٖىٰهَا	ٱللَّهَ
يسٓ	طسٓمٓ	طسٓ	الٓر
قٓ	حمٓ	عٓسٓقٓ	طه
صٓ			
نٓ			

4.2.2 Meem Sakinah:

In this rule, *Meem Sakinah* {مْ} is considered. There are three cases; viz. normal, Hiding, and Merging. If *Meem Sakinah* is succeeded by letter {ب} the rule is Hiding. If succeeded by letter {م} the rule of *Noon* and *Meem Mushadda* is applied (will be described later). In these two cases, a special mark is used with *Meem Sakinah* to denote *Ghunnah*. If succeeded by other than letters {ب، م}, the rule is normal.

4.2.3 Noon and Meem Mushadda:

If any of letters {ن، م} is vocalized by *Shaddah*, it has a nasal tone that is represented a special mark. Similarly, if letter (م) is vocalized by *Shaddah*, it has a nasal tone that is represented by another mark.

4.2.4 Raa:

This rule is concerned about the letter *Raa* {ر}. This letter has mainly two cases; either to be magnified (*Tafkheem*) or softened (*Tarqeeq*). The letter is magnified in any of the following cases:

- It is vocalized by *Fat-ha* or *Dhamma*.
- It has *Sukoon* after *Fat-ha* or *Dhamma*.
- It has *Sukoon* after *Hamzat Wasl*.
- It has *Sukoon* after *Kasra* and followed by a *Tafkheem* letter {خ، ص، ض، غ، ط، ق، ظ}.

On the other hand, the letter *Raa* {ر} is soften in the following cases:

- It is vocalized by a *Kasra*.
- It has a *Sukoon* after a *Kasra* and not followed by a *Tafkheem* letter.
- It is the last letter and preceded by letter *Yaa Saken* {يْ}.

In addition to these two cases, the letter has a special case in one word in which it is between *Fat-ha* and *Kasra*. The word is "مَجْرٖىٰهَا". A special marks are used to denote the three cases.

4.2.5 Qalqala:

Tajweed scholars defined a set of *Qalqala* letters {ق، ط، ب، ج، د} that have special sound when they succeeded by *Sukoon*. These letters have three cases that represent the strength of the *Qalqala*. If any of these letters is the last letter in the given verse, it has two cases: either succeeded by *Shaddah* or *Sukoon*. If succeeded by *Shaddah*, the *Qalqala* case is the strongest. On the other hand, is succeeded by *Sukoon*, the case is a middle case. If *Qalqala* letter is succeeded by *Sukoon* and it is not the last letter in the given verse, it is the weakest case. Special marks are used to represent the *Qalqala* rule and its type.

4.2.6 Compulsory extension:

If any long vowel followed by *Shaddah* or original *Sukoon* – not because of stop – within the same word, there is an extended *Madd*. There is only one word in Quran in which the long vowel "ا" is succeeded by an original *Sukoon* that is "ءَآلْـَٰٔنَ". The other compulsory extended *Madd* is when *Shaddah* succeeds a long vowel. Special marks are used to represent this rule and the type of *Madd*.

4.2.7 Madd with Hamza:

This rule is concerned with long vowels succeeded by letter *Hamza* {ء}. There are two types of this *Madd* depending on the *Hamza* location. Two types of this *Madd* exist; viz. obligatory and optional. If letter *Hamza* is in the beginning of a word, the type is optional. Otherwise, it is obligatory.

4.2.8 Long Vowels:

The long vowels {ـَا, ـُو, ـِي-} are represented as {ا, وو، يي} as the letter *Alif* "ا" does not have any ambiguity and it always comes as a long vowel. However, this is not the case with the other two letters.

4.2.9 General Rule:

If a letter does not satisfy any of the rules above, it is used as it is in the transcription. Table 4 shows Arabic letters used in the transcription as they are.

5. CONCLUSION AND FUTURE WORK

The current work addresses the problem of Quranic text analysis for synthesis of Holy Quran recitation. A technique for transcription and syllabification of Quranic text is proposed. A set of rules is proposed to handle the problems of phonetic transcription and syllabification problems of Quranic text. These rules are set according to Intonation (Tajweed) rules that are set by Quran reciters scholars. The proposed technique can easily be modified to accommodate the problem of transcription and syllabification of standard Arabic text.

The current work has a drawback in the proposed transcription technique as it does not accommodate the sound of the letter follows the nasal tone "Ghunnah". As a future work, we intend to solve the present problem by combining the syllables between which there is a nasal tone and treat them as one syllable. In addition, we intend to make the proposed transcription technique more general by addressing other than Hafss way. Moreover, it is suggested that some information should be added after any verse (Ayah). The added information should address the correct transcription in case of not stopping at the end of a given verse. Finally, the presented work will be employed to synthesize Quranic text. The authors will seek the authentication and approval of specialized scholars and trusted bodies in sciences of the Holy Quran before announcing any system using this work.

Table 3: Syllabified Transcription Examples

#	Quranic Text	Syllabified Phonetic Transcription
1	إِنَّ رَبَّهُم بِهِمْ يَوْمَئِذٍ لَّخَبِيرٌ	ءِنْ5* نَ رَبْ بَ هُم3* بِ هِمْ يَوْ مَ ءِ ذِلْ لَ خَ بِيير ٢
2	قُلْ هُوَ ٱللَّهُ أَحَدٌ	قُلْ هُ وَ ا لْ لَا هُ ءَ حَد$2
3	لَآ أُقْسِمُ بِهَٰذَا ٱلْبَلَدِ	لَا1~ ءُق$3 سِ مُ بِ هَا ذَلْ بَ لَد$2
4	الٓمٓصٓ	ءَ لِفْ لَا3~مْ6* مِيمِ4~صَا4~د$2
5	وَمَآ أَدْرَىٰكَ مَا ٱلْحَآقَّةُ	وَ مَا1~ عَد$3 رَا كَ مَلْ حَا5~قْ قَهْ

Table 4 Arabic Letters Used in the Transcription without Alteration

ز	ذ	خ	ح	ث	ت
ع	ظ	ض	ص	ش	س
		ه	ك	ف	غ

Acknowledgment

The authors would like to acknowledge KFUPM.

References

Abdo, M., Kandil, A., El-Bialy, A., & Fawzy, S. (2010). Automatic detection for some common pronunciation mistakes applied to chosen Quran sounds. 5th Cairo International Biomedical Engineering Conference, (pp. 219–222).

Abdou, S., Hamid, S., Rashwan, M., Samir, A., Abd-elhamid, O., Shahin, M., & Nazih, W. (2006). Computer Aided Pronunciation Learning System Using Speech Recognition Techniques. INTERSPEECH, (pp. 849–852).

Ahmed, A. R. (2004). Quran Phonology: Quran reciting rules based on modern acoustics. Ain Shams University. Master's thesis, Ain Shams University.

Alghamdi, M., Alhamid, A. H., & Aldasuqi, M. M. (2003). Database of Arabic sounds: sentences. Technicat Report, Technicat Report.

Alhuwaykim, M. Z. (2013). Syllabification of single intervocalic consonants in the Arabic dialect of Sakaka City: Evidence from a nonword game. MS Thesis, Southern Illinois University Carbondale, Master's thesis, Southern Illinois University Carbondale.

Alshatebi, A. (2014). Herz Al-Amani Wa Wajh Al-Tahani (حرز الأماني ووجه التهاني). Ghawthani for Quranic Studies.

Azmi, M. M., & Tolba., H. (2008). Syllable-based automatic arabic speech recognition in noisy enviroment. IEEE International Conference on Audio, Language and Image Processing (ICALIP), (pp. 1436-1441).

Eddington, D., Treiman, R., & Elzinga, D. (2013). Syllabification of American English: Evidence from a Large-scale Experiment. Part I∗ . Journal of Quantitative Linguistics, 20(1), 45-67.

Elhadj, Y., Alghamdi, M., & Alkanhal, M. (2013). Approach for Recognizing Allophonic Sounds of the Classical Arabic Based on Quran Recitations. In Theory and Practice of Natural Computing (pp. 57-67). Berlin Heidelberg: Springer.

Elshafei, M., Al-Muhtaseb, H., & Al-Ghamdi, M. (2002). Techniques for high quality Arabic speech synthesis. Information sciences, 140(3), 255-267.

Harrag, A., & Mohamadi, T. (2010). QSDAS: New Quranic Speech Database for Arabic Speaker Recognition. Arabian Journal for Science and Engineering, 35(2), 7-19.

Hernández-Figueroa, Z., Carreras-Riudavets, F. J., & Rodríguez-Rodríguez, G. (2013). Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue. Expert Systems with Applications, 40(17), 7122-7131.

Ibrahim, M. A. (2013). The Shadow Area: A Contrastive Review of the Syllabic Template and Syllabification in English and Fawi Arabic. International Journal of Linguistics, 5(4), 9-21.

Ibrahim, N., Idris, M., Razak, Z., & Rahman, N. A. (2013). Automated tajweed checking rules engine for Quranic learning. Multicultural Education & Technology Journal, 7(4), 275–287.

Imedjdouben, F., & Houacine, A. (2013). Automatic Phonetization of Arabic Text. In Modeling Approaches and Algorithms for Advanced Computer Applications (pp. 85-94). Springer International Publishing.

Imedjdouben, F., & Houacine, A. (2014). Development of an automatic phonetization system for Arabic text-to-speech synthesis. International Journal of Speech Technology, 17(4), 1-10.

Kiparsky, P. (2003). Syllables and moras in Arabic. In The syllable in optimality theory (pp. 147-182). Cambridge University Press.

KSU - Electronic Mosshaf project "Ayat". (2015). Retrieved from http://quran.ksu.edu.sa/

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The PENN arabic treebank: Building a large-scale annotated arabic corpus. NEMLAR conference on Arabic language resources and tools.

Pitakpawatkul, K., Suchato, A., Punyabukkana, P., & Wutiwiwatchai, C. (2013). Thai phonetization of English words using English syllables. 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), (pp. 1 -5).

Qamhawi, A. (1956). Alborhan in Quran Intonation (البرهان في تجويد القرآن).

Ramsay, A., Alsharhan, I., & Ahmed, H. (2014). Generation of a phonetic transcription for modern standard Arabic: A knowledge-based model. Computer Speech & Language, 28(4), 959-978.

Raškinis, A., Raškinis, G., & Kazlauskienė, A. (2003). SAMPA (Speech Assessment Methods Phonetic Alphabet) for encoding transcriptions of Lithuanian speech corpora. Information technology and control, 29(4), 50-56.

Schlippe, T., Ochs, S., & Schultz, T. (2012). Grapheme-to-phoneme model generation for Indo-European languages. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 4801-4804).

Soori, H., Platos, J., Snasel, V., & Abdulla, H. (2011). Simple Rules for Syllabification of Arabic Texts. In Digital Information Processing and Communications (pp. 97-105). Springer Berlin Heidelberg.

The Noble Quran. (n.d.). Retrieved from http://quran.al-islam.org/

Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In Handbook of standards and resources for spoken language systems. Wells, J. C.

Zarrabi-Zadeh, H. (2011). Tanzil Project. Retrieved from http://tanzil.net/wiki/Tanzil_Project