Rule-Based Standard Arabic Phonetization at Phoneme, Allophone, and Syllable Level

Fadi Sindran *fadi.sindran@faui51.informatik.uni-erlangen.de*
Faculty of Engineering
Department of Computer Science /Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Firas Mualla *firas.mualla@cs.fau.de*
Faculty of Engineering
Department of Computer Science /Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Tino Haderlein *Tino.Haderlein@cs.fau.de*
Faculty of Engineering
Department of Computer Science/Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Khaled Daqrouq *haleddaq@yahoo.com*
Department of Electrical and Computer Engineering
King Abdulaziz University,
Jeddah, 22254, Saudi Arabia

Elmar Nöth *noeth@cs.fau.de*
Faculty of Engineering
Department of Computer Science /Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Abstract

Phonetization is the transcription from written text into sounds. It is used in many natural language processing tasks, such as speech processing, speech synthesis, and computer-aided pronunciation assessment. A common phonetization approach is the use of letter-to-sound rules developed by linguists for the transcription from grapheme to sound. In this paper, we address the problem of rule-based phonetization of standard Arabic. [1]The paper contributions can be summarized as follows: 1) Discussion of the transcription rules of standard Arabic which were used in literature on the phonemic and phonetic level. 2) Improvements of existing rules are suggested and new rules are introduced. Moreover, a comprehensive algorithm covering the phenomenon of pharyngealization in standard Arabic is proposed. Finally, the resulting rules set has been tested on large datasets. 3) We present a reliable automatic phonetic transcription of standard Arabic at five levels: phoneme, allophone, syllable, word, and sentence. An encoding which covers all sounds of standard Arabic is proposed, and several pronunciation dictionaries have been automatically generated. These dictionaries have been manually verified yielding an accuracy higher than 99 % for standard Arabic texts that do not contain dates, numbers, acronyms, abbreviations, and special symbols. The dictionaries are available for research purposes.

[1] This paper is an extension of an already published conference contribution [1]. Compared to the conference paper, however, it contains more details and presents new major contributions.

Keywords: Phonetization, Standard Arabic, Phonetic Transcription, Pronunciation Dictionaries, Transcription Rules.

1. INTRODUCTION

Letter-to-sound transcription is the conversion process from the written form to the sound system of a language. It is an essential component in state-of-the-art text-to-speech (TTS), computer-aided pronunciation learning (CAPL), and automatic speech recognition (ASR) systems [2]. The methods of phonetic transcription from grapheme to allophone can be categorized as follows [3]:

- Dictionary-based methods: A lexicon with rich phonological information is utilized to perform the transcription.
- Data-oriented methods: A machine learning-based model is learnt from pronunciation lexicons, and applied then on unseen data.
- Rule-based methods: Expert linguists define letter-to-sound language-dependent rules and a lexicon of exceptions. In these approaches, the degree of difficulty is highly dependent on the extent of compatibility between the writing and sound system of the language as well as the phonetic variations of its sounds.

Arabic, spoken by more than 250 million people, is one of the six official languages of the United Nations [4]. Standard Arabic contains classical Arabic and modern standard Arabic (MSA). Classical Arabic is the language of the Holy Qur'an and books of Arabic heritage. MSA, on the other hand, is the formal language in all Arab countries [5], taught in schools and universities, used in radio and television along with local dialects, and it is the predominant language in which books and newspapers are written. Standard Arabic is known to have a clear correspondence between orthography and the sound system. Therefore, in this work, we adopted a rule-based approach for the transcription. The transcription process of Arabic text can be seen at two levels:

- Phonemic level: The transcription is performed from text to phonemes. Arabic phonemes comprise 28 consonants, 3 short vowels, 3 long vowels, and 2 diphthongs. Table 1 shows the phonemic transcription of Arabic phonemes using the International Phonetic Alphabet (IPA).

consonant	IPA	consonant	IPA	consonant	IPA	consonant	IPA
ء, أ, إ, ؤ, ئ	ʔ	د	d	ض	dˤ	ك	k
ب	b	ذ	ð	ط	tˤ	ل	l
ت, ة	t	ر	r	ظ	ðˤ	م	m
ث	θ	ز	z	ع	ʕ	ن	n
ج	g	س	s	غ	ɣ	ه	h
ح	ħ	ش	ʃ	ف	f	و	w
خ	x	ص	sˤ	ق	q	ي	y
short vowel		**IPA**	**long vowel**		**IPA**	**diphthong**	**IPA**
َ		a	َا, َى		a:	َوْ	aw
ُ		u	ُوْ		u:	َيْ	ay
ِ		i	ِيْ		i:		

TABLE 1: IPA encoding of the Arabic phonemes

- Phonetic level: The effect of neighboring sounds on the phoneme is taken into account. The final produced sound is named allophone or actual phone. Several rules at this level are concerned with the pharyngealization and nasalization of Arabic sounds.

In general, computer-based research on Arabic text and speech, in comparison with English language, is relatively new. Some work done in this domain considered informal languages such as Tunisian Arabic in [5] and Algiers dialect in [6]. Regarding standard Arabic phonetization,

considerable contributions were made by Al-ghamdi [7] [8] and El-Imam [3]. Al-ghamdi et al., in cooperation with Arab linguists, derived transcription rules from Arabic literature books and formulated them in a manner accessible to computer scientists. They implemented these rules in an Arabic TTS system. Moreover, they presented in [8] an encoding system at the phonetic level. El-Imam, on the other hand, in his comprehensive work [3], profoundly analyzed the problems of Arabic text phonetization, letter-to-sound rules, and rule implementation including the assessment of the transcription process outcome. Seen in this context, in this paper, we address the following issues:

- We discuss the above-mentioned works [3], [7], and [8]. In [7], we found that more information is needed in order to apply the rules related to the "ا" /ʔɑlif/ (Alif) correctly. Additionally, some rules at the phonetic level did not consider all the phonemes affected by phonetic variations. Regarding [8], there will be a lack of phonetic information when adopting the encoding they presented. In [3], more details are required about each rule and the priority of rule implementation.
- We propose several modifications to the transcription rules represented in [7]. For instance, the ambiguities concerning the transcription of the Alif have been clarified and the relevant rules have been changed accordingly.
- At the phonetic level, we introduce a comprehensive rule covering the most critical phenomenon at this level, which is the pharyngealization. To the best of our knowledge, there is so far no rule set in the literature which thoroughly and successfully addresses this phenomenon.
- Typically, the performance of a rule-based letter-to-sound approach is heavily dependent on the order of rules. We figure out the correct priority of several rules which is necessary to prevent any undesired overlap.
- The texts used in this work cover all rules at both phonemic and phonetic levels.
- We implemented the transcription rules in a software package to accomplish the following tasks automatically and reliably:
    - Grapheme to phoneme transcription (generation of pronunciation dictionaries). Usually, linguists write these dictionaries for ASR systems manually [9].
    - Grapheme to allophone transcription
    - Grapheme to syllable transcription
    - Statistical analysis at the level of phonemes, allophones, and syllables

At the phonetic level, we modified the encoding developed in [8], so that it can cover all phonetic variations of standard Arabic sounds, and implemented it in the software package. For the pronunciation dictionaries, we adopted the Arpabet coding [10], which is a phonemic transcription system that depends on English capital letters and digits developed by the Advanced Research Projects Agency (ARPA). Therefore, the output of the transcription can be utilized in any Arpabet-compatible system, such as the TTS and ASR tools in the widely-used CMU Sphinx Toolkit.

2. PRE-TRANSCRIPTION

In this work, we assume that the input of the transcription process is an Arabic text which is fully diacritized. Additionally, before applying the transcription rules, the following steps need to be performed:

- Cleaning the text by keeping only Arabic characters and symbols with a single sentence per line.
- Converting dates, numbers, acronyms, abbreviations, and special symbols to a proper word or sequence of words manually.
- Words of irregular spelling, i. e. words which do not follow the Arabic pronunciation rules, are processed separately. We collected most of these words in a separate lexicon which can be extended later to include new words.

3. PHONEMIC LETTER-TO-SOUND RULES

In our work, we adopted the rules proposed by Al-ghamdi et al. in [7] after adjusting them. For the phonemic level, these rules are listed in Section 3.1 while our modifications are presented in Section 3.2.

3.1. State-of-the-art Rules At The Phonemic Level

Following [7], phonemic transcription is achieved according to the following rules (in order):

1) Convert geminated consonants (consonants with a diacritic named Shaddah written as "ّ") to two consecutive ones.
2) Delete the grapheme "ا" /ʔɑlif/ (Alif) if it is followed by two consecutive consonants and does not occur at the beginning of a phrase.
3) If the Alif is a part of the "ال" (the definite article in Arabic, IPA: /ʔɑlluttɑʕri:f/) at the beginning of a phrase, it is pronounced as "ء" /ʔɑ/.
4) If the Alif is a part of a verb, whose third letter is diacritized with an original "ُ" /dˤɑmmah/ (Dammah), and appears at the beginning of a phrase, it should be converted to "ء" /ʔu/, otherwise convert the Alif to "ء" /ʔi/ when it appears at the beginning of a phrase.
5) Convert an Alif preceded with a "َ" /fatħah/ (Fatha) to a long Fatha "ا" /ɑ:/.
6) Convert a "و" /wɑ:w/ preceded by a "ُ" and followed by a Sukun (diacritic written as "ْ") to a long Dammah "وُ" /u:/.
7) Convert a "ي" /yɑ:ʔ/ preceded by a "ِ" /i/ (Kasrah) and followed by a Sukun to a long Kasrah "يِ" /i:/.
8) Convert two similar consecutive short vowels to one long vowel.
9) Convert a "ة" /tɑ:ʔ mɑrbutˤɑh/ to "ه" /hɑ:ʔ/ when it occurs at the end of a phrase.
10) Delete the shamsi "ل" /lɑ:m/. For example, the transcription of the word "النُّور" /ʔɑlnnu:r/ (the light) is "أَنُّور" /ʔɑnnu:r/.
11) Delete the Tanwin (a diacritic which may have one of three forms: "ً", "ٌ", or "ٍ") at the end of a phrase and turn it to short vowel and "ن" /n/ anywhere else.
12) Convert "آ" /ʔɑlifulmɑd/ to "اء" /ʔɑ:/ wherever it is found in the text.
13) Pronunciation of all types of the Hamza (the types are: "ء", "أ", "إ", "ؤ", and "ئ") is "ء".
14) When three consonants occur consecutively, a short vowel has to be inserted after the first one. This inserted short vowel is either "َ", "ُ", or "ِ", given that the short vowel which precedes the first consonant is "ِ", "ُ", or "َ", respectively.

3.2. Discussion and Proposed Modifications At The Phonemic Level

In rule "2", we must distinguish between the long vowel "ا", as in "ضَالِّين" /dˤɑ:lli:n/ (erratic), and "ا" (Arabic: "همزة الوصل" /hɑmzɑtulwasˤl/), as in "فَاتَّبَاع" /fɑttibɑ:ʕ/ (following): Only in the case of Hamzat al-Wasl (the second case), the Alif must be deleted when it is followed by two consonants. In order to solve this problem, we made:

- a list L_1 of Arabic words which have Hamzat al-Wasl followed by a geminated consonant with all possible prefixes. We extracted these words from the famous Arabic dictionary Lisan Al Arab (the tongue of the Arabs, Arabic: "لسان العرب") [11].
- a list L_2 of proper names which have a long vowel "ا" followed by two consonants, such as "إِيطَالِيَا" /ʔitɑ:lyɑ:/ (Italy) and "أَلْمَانِيَا" /ʔɑlmɑ:nyɑ:/ (Germany).

The application of rule "1" and rule "2" in Section 3.1 is then modified as described in Algorithm 1. For the sake of clarity, the pseudocode texts of all algorithms in this paper are annotated with examples.

```
1:  if a word W ∈ L1 then
2:    rule "2" is applied
3:    rule "1" is applied # e. g. "وَاتِّهَام" /wɑttihɑ:m/ (accusation)
4:  end if
5:  if a word W ∈ L2 then
6:    rule "1" is not applied
7:    rule "2" is not applied # e. g. "رُومَانيَا" /ru:mɑ:nyɑ:/ (Romania)
8:  end if
9:  if a word W ∉ (L1&L2) then
10:   rule "1" is applied
11:   rule "2" is not applied # e. g. "جَادِّين" /gɑ:ddi:n/ (earnest)
12: end if
```

ALGORITHM 1: rules "1" and "2" after modification.

In rule "3", we cannot distinguish whether the "ال" is a definite article or not. For example, the "ال" in the word "البَرد" /ʔɑlbɑrd/ (cold) is a definite article and must be converted to "ءَل" /ʔɑl/, but the "ال" in the word "التِمَاس" /ʔiltimɑ:s/ (request) is not a definite article. In fact, contrary to rule "3", it must be converted to "ءِل" /ʔil/. Moreover, rule "4" is not valid with verbs which end with "ا" or "ى" /ʔɑlif mɑqsˤu:rɑh/ in the past tense. The first Alif here must be converted to "ءِ" /ʔi/ regardless of the diacritic at the fourth position for instance in the verb "ارمُوا" /ʔirmu:/ (throw) where the first "ا" must be always converted to "ءِ" /ʔi/. We suggest to modify rule "3" and rule "4" as described in Algorithm 2.

```
1:  if the third letter l3, in a word starting with "ال", is a shamsi letter (1) then
2:    if l3 is followed by "ّ" or "ِّ" then
3:    convert "ال" to "ءِل" /ʔil/
      # e. g. "التَهَمَ" /ʔiltɑhɑmɑ/ (engorge),
      # "التِهَاب" /ʔiltihɑ:b/ (inflammation)
4:    else if l3 is followed by "ُّ" then
5:    convert "ال" to "ءُل" /ʔul/
      # e. g. "التُمِسَ" /ʔultumisɑ/ (it was petitioned)
6:    end if
7:  end if
8:  if a verb ends with "ا" or "ى" in the past tense (2) then
9:     rule "4" may not be applied
10:   the Alif must be converted to "ءِ" /ʔi/
      # e. g. "سَمَا" /sɑmɑ:/ (fly), "مَشَى" /mɑʃɑ:/ (walk)
11: end if
12:   convert "ال" in all other words to "ءَل" /ʔɑl/
      # e. g. "البَرق" /ʔɑlbɑrq/ (lightning)
      # note (1): shamsi letters in standard Arabic are: "ت", "ث", "د", "ذ", "ر",
      # "ز", "س", "ش", "ص", "ض", "ط", "ظ", "ل", and "ن".
      # note (2): we gathered all verbs which end with "ا" or "ى" in the past tense
      # in a separate lexicon.
```

ALGORITHM 2: rules "3"and "4" after modification.

Lastly, in rule "14", in the case when the first consonant is "ن" resulting after the transcription of the Tanwin (see rule "11"), a modification needs to be made. In this case, the short vowel which must be added is "ِ" regardless of the short vowel before the first consonant. For example, the transcription of the sentence "إِنَّهُ خَالِدٌ السَّمَّان" (he is Khalid the grocer) is "إِنَّهُ خَالِدُنِ سَّمَّان" /ʔinnɑhuxɑ:lidunissammɑ:n/. Here, we add "ِ" after "ن" in the word "خَالِدُن" in spite of the fact that the short vowel before "ن" is "ُ".

4. SYLLABICATION

At this point, the phonemic content of the text is available as an output of the phonemic transcription process described in Section 3. It is thus possible to aggregate the resulting phonemes in higher linguistic units, i. e. syllables. Syllabication is very important in Arabic TTS systems because stress depends on the type of syllables in the word [12]. Recognition of non-Arabic words as well as learning melody and meters of Arabic poetry are other application areas of syllabication. Moreover, as will become clear in Section 5.2.2, the pharyngealization rules in Arabic are syllable-based. The Arabic syllables can be characterized as follows:

- Every syllable must start with a consonant followed by a vowel.
- The number of vowels in a word is equal to the number of syllables in this word.

There are six types of syllables in Arabic: CV, CV:, CVC, CV:C, CVCC, and CV:CC, where C refers to a consonant, V to a short vowel, and V: to a long vowel. In order to represent the two diphthongs, we categorized the syllables into eight types instead of six as follows: CV, CD2, CL, CVC, CD2C, CLC, CVCC, and CLCC, where L refers to a long vowel and D2 is a diphthong. We use the procedure described in Algorithm 3 for the syllabication. In this algorithm, "***P***" denotes the phonemic transcription of a given text. Moreover, "***S***" denotes the transcription we obtain after replacing consonants, short vowels, long vowels, and diphthongs in "***P***" with the symbols C, V, L, and D2, respectively.

1: find the number of syllables $\mathbf{N_s}$ which is the total number of occurrences of the symbols
   V, **L**, and **D2** in ***S***
2: **if** $\mathbf{N_s} > 1$ **then**
   while ***l***(***S***) > 3 **do** # ***l***(***S***): the length of ***S***
4: **if** the fourth symbol in ***S*** is **C** **then**
5: take the first three phonemes in ***P*** as one syllable
6: take the first three symbols in ***S*** as the type of this syllable
7: remove the first three phonemes or symbols from ***P*** and ***S***
8: **else**
9: take the first two phonemes in ***P*** as one syllable
10: take the first two symbols in ***S*** as the type of this syllable
11: remove the first two phonemes or symbols from ***P*** and ***S***
12: **end if**
13: **end while**
14: take the phonemes in ***P*** as one syllable
15: take the symbols in ***S*** as the type of this syllable (1)
16: **else**
17: do (1)
18: **end if**

ALGORITHM 3: Syllabication algorithm.

Table 2 shows an example in which Algorithm 3 is employed to syllabify the word "السَّلَام" /ʔɑssɑlɑːm/ (peace).

P	ء	ـَ	س	س	ـَ	ل	ـَا	م
***P* (Arpabet)**	E	AE	S	S	AE	L	AE:	M
S	C	V	C	C	V	C	L	C
1st iteration	C	V	C	**C**				
1st syllable	E	AE	S					
2nd iteration				C	V	C	**L**	
2nd syllable				S	AE			
3rd iteration						C	L	C
3rd syllable						L	AE:	M

TABLE 2: Syllabication of the word "السَّلَام" using Algorithm 3.

5. PHONETIC LETTER-TO-SOUND RULES

5.1. State-of-the-art Rules At The Phonetic Level

The rules at this level are used for transcription from phoneme to allophone. Following [7], they are given in the order of application as follows:

1) Convert "ن" /n/ followed by "ب" /b/ or "م" /m/ to "م".
2) Convert "ن" followed by "ر" /r/ to "ر".
3) Convert "ن" followed by "ل" /l/ to "ل".
4) Convert "ن" to "و" /w/ or "ي" /y/ with nasalization, when a word ends with "ن" and the next one starts with "و" or "ي", respectively.
5) Convert "ن" to a nasal sound of the following letter when it is followed by any letter except "و" or "ي".
6) Convert "ذ" /ð/ followed by "ظ" /ðˤ/ to "ظ".
7) Convert "ت" /t/ followed by "ط" /tˤ/ to "ط".
8) Convert "ت" followed by "د" /d/ to "د".
9) Convert "د" followed by "ت" to "ت" when they occur in one word.
10) Convert "د" to "ت" when "ت" occurs after the word "قَد" /qɑd/ (may).
11) Convert "ل" followed by "ر" to "ر".
12) The "ل" and "ر" are pronounced sometimes normally (light accent, in Arabic: "تَرقِيق" /tɑrq i:q/) and other times with pharyngealization (heavy accent, in Arabic: "تَفخِيم" /tɑfxi:m/) depending on the context in which they occur. Further details can be checked in [7].
13) The vowels followed by "قك" /qk/ or "طت" /tˤt/ are pronounced with a heavy accent.
14) Arabic vowels followed by emphatic consonants ("خ" /x/, "غ" /ɣ/, "ق" /q/, "ص" /sˤ/, "ض" /dˤ/, "ط" /tˤ/, "ظ" /ðˤ/), pharyngeal "ل", or pharyngeal "ر" are pharyngealized.
15) Make short vowels shorter at the end of a word followed by another one, or when they are followed by two consonants.
16) Convert a long vowel followed by two consonants to a short one.
17) When stopping at the end of a phrase, the short vowels must be deleted.
18) Release stop sounds at the end of a phrase (Arabic: "قَلقَلَة" /qɑlqɑlɑh/, with voiced stops ("ق", "ط", "ب", "ج" /g/, and "د"), and aspiration with voiceless stops ("ء" /ʔ/, "ت", and "ك" /k/).

5.2. Discussion and Proposed Modifications At The Phonetic Level

We suggest, in Section 5.2.1, to extend the aforementioned set of phonetic-level rules using two additional allophones. In Section 5.2.2, we profoundly address the problem of pharyngealization in standard Arabic and propose a comprehensive set of rules that can deal with the complications of this phenomenon.

5.2.1. Additional Allophones

There is one allophone for each of the short vowels "ُ" /u/ and "ِ" /i/ that must be taken into account in the phonetization process. These allophones occur in the syllables of type CVC, when stopping at the end of a word (e.g. at the end of the sentence), or when a word consists only of one syllable of this type. For instance, in the word "جُندُب" /gundub/ (grasshopper) which consists of two syllables "جُن" /gun/ and "دُب" /dub/, the short vowel "ُ" in the second syllable is pronounced differently from the short vowel in the first syllable. For a correct pronunciation of this word, during the first vowel, the tongue has to be at a high position (close to the roof of the mouth), while it has to be low during the second one. As mentioned above, this phenomenon is also observed in the words which consist of a single CVC syllable. An example of this case is the word "جُد" /gud/ (be generous), where the short vowel "ُ" is the same as the one in the syllable "دُب" from the previous example. There is an exception to this rule in the words "مِن" /min/ (of), "إذ" /ʔið/ (as), and "إن" /ʔin/ (if). The sound "ِ" in this case has no allophone. Other examples that contain the aforementioned additional allophones can be found in Table 3.

word	syllables		tongue position	
قُنفُذ /qunfuð /	فُذ /fuð/	قُن /qun/	low	high
قُل /qul/	قُل /qul/		low	
هُدهُد /hudhud/	هُد /hud/	هُد /hud/	low	high
سِمسِم /simsim/	سِم /sim/	سِم /sim/	low	high
ارمِه /ʔirmih/	مِه /mih/	ءِر /ʔir/	low	high

TABLE 3: Examples of the additional allophones for "ُ" and "ِ".

5.2.2. Pharyngealization

Considering the relevant literature, one can notice that there is no complete agreement about the positions of pharyngealization and affected sounds in Arabic. Al-ghamdi et al. in [7] considered only the vowels that followed the pharyngealization sounds. This is, however, not sufficient as the two diphthongs, the sounds /ت/, /د/, /ذ/, /س/, and the non-pharyngealized /ل/ and /ر/ must be considered as well. In fact, pharyngealization may affect some sounds before the pharyngealized sound. For example, in the word "مُستَطِيل" /mustaʈˤi:l/ (rectangle), all the sounds /ُ/, /س/, /ت/, and /َ/ before /ط/ are pharyngealized. Abu Salim in [13] considered only the sounds which are in the same syllable where the pharyngealization sound occurs. He ignored the sounds in the syllables which come before or after. However, these sounds can be affected as well. For instance, the sound /د/ in the word "مِنضَدَة" /mindˤadah/ (table) is pharyngealized even though it does not occur in the syllable which contains the pharyngealized sound /ض/. El-Imam in [3] presented an important discussion of the pharyngealization of Arabic sounds, but he did not consider the different types of syllables. We think that the phenomenon of pharyngealization in Arabic is more complex than what is described in [3], [7], and [13]. In the text which follows, we propose a general rule for pharyngealization in standard Arabic. This rule is a result of profound analysis of a representative sample of Arabic sentences pronounced and analyzed by native speakers.

A comprehensive rule of pharyngealization in standard Arabic

- The sounds causing pharyngealization in standard Arabic can be divided into two sets:
  - Set_1: contains /ص/, /ض/, /ط/, and /ظ/, in addition to the sounds /س/, /د/, /ت/, and /ذ/ when they are affected by the aforementioned sounds. The mechanism of this effect will be described later in Algorithm 4.
  - Set_2: contains /خ/, /غ/, /ق/, pharyngealized /ل/, and pharyngealized /ر/.

- The pharyngealization caused by Set_1 depends on the syllables, their types, and the positions of the pharyngealized and non-pharyngealized sounds with respect to the syllable which contains a sound from Set_1.
- The sounds which belong to Set_1 affect the short vowels except "ِ", long vowels except "ِي", "ت", "د", "ذ", "س", and the two diphthongs.
- The Set_1-type pharyngealization is described in Algorithm 4. In this algorithm, S^0 refers to the syllable which contains a sound from Set_1, S^{-1} refers to the syllable that precedes S^0 directly, whereas S^{+1} refers to the syllable that follows S^0 directly. It is necessary pointing out that the order of rule application in the algorithm is critical for the validity of the results.

1: **if** S^0 belongs to a single word W^0 **then**
2: apply pharyngealization to the sounds in S^0 (see the text for the sounds affected by Set_1-type pharyngealization).
3: **switch** type of S^0 **do**
4: **case CVC**
5: **if** S^{-1} belongs to W^0 and contains "ت", "د", "ذ", or "س" **then**
6: pharyngealization affects the sounds in S^{-1} (1)
e.g. "وَسِيطٌ" /Wɑ si: tˤun/.
7: **end if**
8: **if** S^{-1} belongs to W^0 and does not have "ِ" or "ِي" **then**
9: pharyngealization affects the sounds in S^{-1} (2)
e.g. "وَضَربٌ" /wɑ dˤɑr bun/.
10: **end if**
11: **if** the sound at the third position of S^0 belongs to Set_1 **then**
12: **if** S^{+1} belongs to W^0 and contains "ت", "د", "ذ", or "س" **then**
13: pharyngealization affects the sounds in S^{+1} (3)
e.g. "قِصدِير" /qisˤ diːr/.
14: **end if**
15: **if** S^{+1} belongs to W^0 and does not contain /ِ/ or /ِي/ **then**
16: pharyngealization affects the sounds in S^{+1} (4)
e.g. "مَصنَعٌ" /mɑsˤ nɑ ʕun/.
17: **end if**
18: **if** the short vowel in S^0 is not /ِ/ **then**
19: apply (3), (4)
e.g. "قَطرَةٌ" /qɑtˤ rɑ tun/.
20: **end if**
21: **end if**
22: **case CV**
23: apply (1), (2), (3), and (4)
e.g. "وَضَرَبَ" /Wɑ dˤɑ rɑ bɑ/, "طَبَرقَة" /tˤɑ bɑr qɑh/.
24: **case CL**
25: **if** the long vowel in S^0 is /ِي/ **then**
26: apply (1), (2)
e. g. "وَصِيفَة" /wɑ sˤiː fɑh/.
27: **else**
28: apply (1), (2), (3), and (4)
e. g. "وَطَافَ" /wɑ tˤɑː fɑ/.
29: **end if**
30: **case CD2**, **CD2C**, **CVCC**, or **CLCC**
31: apply (1), (2)
e. g. "وَضَيفٌ" /wɑ dˤay fun/, "وَصَيف" /wɑ sˤɑyf/,
"وَالأَرض" /wɑl ʔɑrdˤ/, "وَضَارّ" /wɑ dˤɑːrr/.
32: **case CLC**
33: **if** the sound of pharyngealization is only at the first position **then**
34: do as in the **case CL**
e. g. "وَضَارٌّ" /wɑ dˤɑːr run/.
35: **end if**
36: **if** the sound of pharyngealization is at the third position,
37: or at both the first and third positions **then**
38: apply (1), (2)
e. g. "وَفَاض" /wɑ fɑːdˤ/.
39: **end if**
40: **else**
41: **if** the type of S^0 is **CVC then**
42: **if** the sound of pharyngealization in S^0 is only at the first position **then**
43: **if** S^0 does not contain /س/, /ذ/, /د/, /ت/, or /ل/ **then**
44: apply pharyngealization to the sounds in S^0.
e. g. "بَعضُ الزَّمَر" /bɑʕ dˤuz zu mɑr/.
45: **end if**
46: apply (1) and (2)
e. g. "حَوضُ البَحر" /ħɑw dˤul bɑħr/.

```
47:     else if the sound of pharyngealization in S^0 is only at the third position then
48:         apply pharyngealization to the sounds in S^0.
            # e. g. "أَشرَقَ الصُّبح" /ʔaʃ ra qasˤ sˤubħ/.
49:         apply (3) and (4)
50:     else
51:         apply pharyngealization to the sounds in S^0
52:         apply (1), (2), (3), and (4)
            # e. g. "بَعدَ الظُّهر" /baʕ daðˤ ðˤuhr/.
53:     end if
54:   end if
55: end if
```

ALGORITHM 4: Set_1-type pharyngealization.

Table 4 shows more examples of the Set_1-type pharyngealization.

word or phrase	syllables	syllable types	pharyngealization yes (y) or no (n)
مَن صَادَقَهُ /mansˤaːdaqahu/	مَن صَا دَ قَ هُ	CV CV CV CL CVC	n y y y n
وَصَاحَ /wasˤaːħa/	وَ صَا حَ	CV CL CV	y y y
طَارق /tˤaːriq/	طَا رق	CVC CL	n y
صَولَجَان /sˤawlagaːn/	صَو لَ جَان	CLC CV CD2	n n y
ظَافِر /ðˤaːfir/	ظَا فِر	CVC CL	y y
اِستِطَالَة /ʔistitˤaːlah/	ءِس تِ طَا لَه	CVC CL CV CVC	y y y y
طَبَرقَة /tˤabarqah/	طَ بَر قَه	CVC CVC CV	y y y
مَصرَع /masˤraʕ/	مَص رَع	CVC CVC	y y
حَافِظ /ħaːfiðˤ/	حَا فِظ	CVC CL	y y
بِضعَة /bidˤʕatun/	بِض عَ تُن	CVC CV CVC	n y y
وَالأرض /walʔardˤ/	وَل ءَرض	CVCC CVC	y y
وَالفِظ /walfiðˤ/	وَل فِظ	CVC CVC	y y
عِندَ الأيض/ʕindalʔaydˤ/	عِن دَل ءَيض	CD2C CVC CVC	y n n
صَعبٌ /sˤaʕbun/	صَع بُن	CVC CVC	y y
ضَالِّين /dˤaːlliːn/	ضَال لِين	CLC CLC	n y
وَضَرب /wadˤarb/	وَ ضَرب	CVCC CV	y y
وَضَارّ /wadˤaːrr/	وَ ضَارر	CLCC CV	y y
وَضَيفٌ /wadˤayfun/	وَ ضَي فُن	CVC CD2 CV	n y y
وَصَيف /wasˤayf/	وَ صَيف	CD2C CV	y y
وَصِيفَة /wasˤiːfah/	وَ صِي فَه	CVC CL CV	n y y
طِينَة /tˤiːnatu/	طِي نَ ةُ	CV CV CL	n n y
مَنَاظِيرٌ /manaːðˤiːrun/	مَ نَا ظِي رُن	CVC CL CL CV	y y y n
وَضِيعٌ /wadˤiːʕun/	وَ ضِي عُن	CVC CL CV	n y y

TABLE 4: Examples of the Set_1-type pharyngealization.

- In the words "اللَّه"/ʔallaːh/ (God) and "اللَّهُمَّ"/ʔallaːhummaˌ/ (o Allah), with all possible prefixes, the pharyngealized /ل/ affects the short vowel /ـَ/ or /ـُ/ before it, the Alif /ا/, and the short vowel /ـَ/ or /ـُ/ after the sound /ه/. In the word "أَلَّه" /ʔaːllaːh/, the long vowel /ا/ before /ل/ is pharyngealized. Table 5 illustrates this rule.

word	syllables			Pharyngealization yes (y) or no (n)		
وَاللَّهُ /wallaːhu/	هُ /hu/	لَا /laː/	وَل /Wal/	y	y	y
وَاللَّهَ /wallaːha/	هَ /ha/	لَا /laː/	وَل /Wal/	y	y	y
أَلَّهُ/ʔaːllaːhu/	هُ /hu/	لَا /laː/	ءَال /ʔaːl/	y	y	y

TABLE 5: Pharyngealization induced by pharyngealized /ل/.

5.3. Phonetic Level Encoding

The result of the transcription is represented after [8] (cf. Table 6) using two letters to represent the phoneme, a number to represent sound duration including geminates, and another number to represent the phonetic variations.

position	value interpretation		
1, 2	two letters represent a phoneme		
3 (gemination)	1 (not geminated)	2 (geminated)	
4 (ph. var.)	4 (nasalized)	5 (released with a schwa)	6 (centralized)

TABLE 6: Encoding of the transcription result after [8] (ph. var.: phonetic variations).

We extended this encoding in order to cover all phonetic variations of Arabic phonemes. Our encoding consists of two characters and three digits instead of two characters and two digits. The first two characters are two Latin small letters representing the phoneme as in the previous encoding, the first digit represents the pharyngealization feature, the second digit symbolizes sound duration, and the last digit represents all phonetic variations except the features related to pharyngealization. Our modifications to the encoding are in the two digits we used to cover the allophones that are not observed in the encoding developed by Alghamdi in [8]. These allophones are related to the values "0", and "1" of the first digit, and "3", "4", "5", "6", "7", "8", and "9" of the last digit in the encoding we present in Table 7.

position	value interpretation		
1, 2	two letters represent a phoneme		
3 (pharyngealization)	0 (not pharyngealized)	1 (pharyngealized)	
4 (duration)	0 (short)	1 (shorter)	2 (long)
5 (phonetic variations)	0 (the sound is a phoneme)	1 (allophone for "ن")	
	2 (released with a schwa)	3 (allophone for "ض" followed by "ط")	
	4 (allophone for "ت" followed by "د")	5 (allophone for "د" followed by "ت")	
	6 (allophone for "ل" followed by "ر")	7 (allophone for "ق" followed by "ك")	
	8 (allophone for "ط" followed by "ت")	9 (allophone for /ُ/ or /ِ/)	

TABLE 7: The modified encoding at phonetic level.

Table 8 shows some examples of the modified encoding. Note that the two additional allophones suggested in Section 5.2.1 are represented by the value 9 at the fifth position.

text	شَكَرنَاهُم /ʃɑKɑrnɑ:hum/ (we thanked them)
encoding	js000 as000 ks000 as110 rs100 ns000 as020 hs000 us009 ms000
text	مِن فَضلِك /minfadˤlik/ (please)
encoding	ms000 is010 fs001 fs000 as110 db000 ls000 is009 ks002
text	أينَ الكِتَاب /ʔaynalkita:b/ (where is the book?)
encoding	hz000 ay000 ns000 as010 ls000 ks000 is000 ts000 as020 bs002
text	جَاءَ التِّلمِيذُ مِن مَدرَسَتِه /ga:ʔattilmi:ðuminmadrasatih/ (the pupil came from his school)
encoding	jb000 as020 hz000 as010 ts000 ts000 is010 ls000 ms000 is020 vb000 us010 ms000 is010 ms001 ms000 as010 ds000 rs100 as100 ss000 as000 ts000 is009 hs000

TABLE 8: Examples of the modified encoding at the phonetic level.

6. CORPORA USED IN THIS WORK

The corpora used in this work can be divided into two sets:

- Set_a: contains two corpora that have been used to find new rules and/or new details of existing rules as well as to figure out the correct priority of all rules. The set_a corpora are:

- KACST corpus: Developed by King Abdul-Aziz City for Science and Technology (KACST) and contains 367 fully diacritized sentences including altogether 1812 words [14]. This corpus is characterized as follows:
  - ✓ Minimum repetition of words in all sentences
  - ✓ Every sentence contains between two and nine words.
  - ✓ Sentences can be easily pronounced.
  - ✓ The sentences were chosen in a way which minimizes the number of sentences and maximizes the phonetical information content. This corpus was used in a project of KACST and IBM for training a speech recognizer for Arabic words.

- Holy Qur'an corpus: The holy Qur'an [15] without Surat al-Baqarah /suratulbaqarah/ (Arabic: "سورة البقرة"). This corpus contains 72146 words.

- Set_b: contains five corpora for testing:
  - Surat al-Baqarah corpus: Surat al-Baqarah is the longest sura of the Qur'an containing 286 verses.
  - Sahih al-Bukhari corpus: Vocabulary of Sahih al-Bukhari /sˤaħi:ħulbuxa:ri:/ (Arabic: "صحيح البخاري"), the book that contains the most correct prophetic traditions [16].
  - Nahj al-Balagha corpus: A set of sentences selected from Nahj al-Balagha (Way of Eloquence), which is a book that contains sermons, letters, and sayings of Imam Ali ibn Abi Talib [17].
  - Umm Alqura corpus: A list of most frequently used Arabic words [18], prepared by Institute of Arabic Studies in the University of Umm Alqura. This list contains 5446 words.
  - Katharina Bobzin corpus: Made by Katharina Bobzin from Erlangen university, it contains 306 sentences selected from her book: "Arabisch Grundkurs" (English: Arabic Basic Course) [19].

Additional information and statistics about these corpora are given in tables 9, 10.

corpus	KACST corpus	Holy Qur'an corpus
vocabulary	1413	16675
phonemes/allophones	11134	453342
syllables	4533	194901

TABLE 9: Information and statistics about set_a corpora.

corpus	Surat al-Baqarah corpus	Sahih al-Bukhari corpus	Nahj al-Balagha corpus	Umm Alqura corpus	Katharina Bobzin corpus
non-existent words in KACST corpus (%)	95.48	98.82	96.30	98.06	90.84
non-existent words in Holy Qur'an corpus (%)	68.21	82.29	86.57	88.18	84.64
vocabulary	2542	50468	3898	5044	742
phonemes/allophones	38827	422530	44901	28762	8224
syllables	16710	185321	19424	10735	3451

TABLE 10: Information and statistics about set_b corpora.

7. EVALUATION

The transcription of a word depends on its location in the sentence. We thus used both words and sentences to test the transcription rules, the transitions between words are taken into account in the evaluation. We implemented our transcription algorithm in Matlab. The output of the program includes:

- Phonemic dictionary of the entire input text given in Arpabet.

- Phonetic dictionary represented according to the encoding proposed in Section 5.3.
- Syllables transcription including syllable types.

The automatically generated results from the five test corpora (cf. Section 6) were validated manually by expert linguists. We checked the results at four levels: phoneme, allophone, syllable, and word. The achieved accuracy at each level is given in Table 11, while the words, where transcription errors occurred, are presented in Table 12.

test corpora	Surat al-Baqarah corpus	Sahih al-Bukhari corpus	Nahj al-Balagha corpus	Umm Alqura corpus	Katharina Bobzin corpus
phoneme accuracy (%)	100	100	100	99.98	99.99
allophone accuracy (%)	100	100	100	99.98	99.99
syllable accuracy (%)	100	100	100	99.94	99.97
word accuracy (%)	100	100	100	99.88	99.93

TABLE 11: Test results.

Word	**Correct phonemic transcription (IPA)**	**phonemic transcription according to this work**
يُوليُو (July)	yuːlyuː	yulyuː
يُونيُو (June)	yuːnyuː	yunyuː
فَرَاولَة (strawberries)	fɑrɑːwlɑh	fɑrɑwlɑh
جُغرَافيَا (geography)	ɡuɣrɑːfyɑː	ɡuɣrɑfyɑː
فِيزيَاء (physics)	fiːzyɑːʔ	fizyɑːʔ
كِيمِيَاء (chemistry)	kiːmyɑːʔ	kimyɑːʔ
التَّاكسِي (the taxi)	ʔɑttɑːksiː	ʔɑttɑksiː

TABLE 12: Words with transcription errors after the evaluation.

We obtained seven errors in seven words where the long vowels /ɑː/, /uː/, and /iː/ in the phonemic transcription have been replaced by the short corresponding ones /ɑ/, /u/, and /i/, respectively. Obtaining errors in the words presented in Table 12 is expected since they are non-Arabic words. We could not compare our results with [7] because Al-ghamdi cited only handmade examples to test the rules. In order to compare our results with El-Imam in [3], we used the same list of common vocabulary [18] he used after rewriting the 5044 words in this list in a text file manually, as this list is not available for direct use. We considered only the errors caused by grapheme-to-phoneme and phoneme-to-phone conversion. The list of sentences used by El-Imam was not available. Table 13 shows the comparison result at the level of phonemes, allophones, and words.

Corresponding work	El-Imam [3]	this work
Phoneme and allophone error rate (%)	1.13	0.02
Word error rate (%)	4.54	0.12

TABLE 13: Comparison result for the data used by El-Imam in [3].

From Table 13 we find that our work yields an improvement with the automatic transcription of about 1 percent point at the level of phoneme and allophone altogether and about 4 percent points at the level of words.

8. CONCLUSION

We tackled the problem of rule-based automatic phonetization of Arabic text. Our method was tested on five corpora representing rich phonetical information and containing together more than 54500 vocabulary words. The results show that the proposed method yields an accuracy higher than 99% at four levels: phoneme, allophone, syllable, and word. In addition to words, complete sentences were used in the evaluation so that the transitions between the words are considered.

We generated pronunciation dictionaries from these corpora automatically and validated the correctness of the results manually. A profound discussion of the rule-based Arabic phonetization was presented, and several extensions and improvements to the state-of-the-art were proposed. A major contribution of this paper is the proposed set of rules which addresses the pharyngealization in standard Arabic. This phenomenon is subtle and difficult to model in a rule-based manner even for native speakers. Compared to the previous version of our method as presented in [1], two additional corpora of more than 50000 vocabulary words have been used for testing in this work. As future work, we consider the automatic processing of acronyms, abbreviations, special symbols, numbers, proper names, and words with irregular spelling. Moreover, we are currently working on an Arabic TTS engine which builds on the results presented in this study.

9. REFERENCES

[1] F. Sindran, F. Mualla, K. Bobzin, E. Nöth, “Automatic robust rule-based phonetization of standard Arabic,” in: Text, Speech, and Dialogue, Vol. 9302 of LNAI, Springer, 2015, pp. 442–451.

[2] M. Ali, M. Elshafei, M. Al-Ghamdi, H. Al-Muhtaseb, A. Al-Najjar, “Arabic phonetic dictionaries for speech recognition,” Journal of Information Technology Research 2, 2009, pp. 67–80.

[3] Y. El-Imam, “Phonetization of arabic: rules and algorithms,” Computer Speech & Language 18, 2004, pp. 339–373.

[4] K. Hadjar, R. Ingold, “Arabic newspaper page segmentation,” in: 7th International Conference on Document Analysis and Recognition, Vol. 2, 2003, pp. 895–899.

[5] A. Masmoudi, M. Ellouze Khemakhem, Y. Estève, L. Hadrich Belguith, N. Habash, “A corpus and phonetic dictionary for tunisian arabic speech recognition,” in: LREC, 2014, pp. 306–310.

[6] S. Harrat, K. Meftouh, M. Abbas, K. Smaili, “ Grapheme to phoneme conversion: an arabic dialect case,” in: 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU’14), 2014.

[7] M. Al-ghamdi, H. Al-Muhtasib, M. Elshafei, “Phonetic rules in arabic script,” Journal of King Saud University - Computer and Information Sciences 16, 2004, pp. 85–115.

[8] M. Alghamdi, Y. O. M. El Hadj, M. Alkanhal, “A manual system to segment and transcribe arabic speech,” in: IEEE International Conference on Signal Processing and Communications (ICSPC), 2007, pp. 233–236.

[9] F. Biadsy, N. Habash, J. Hirschberg, “Improving the arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules,” in: Proceedings of Human Language Technologies, The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL/HLT), 2009, pp. 397–405.

[10] Arpabet, Internet: https://en.wikipedia.org/wiki/Arpabet [October 23, 2016].

[11] I. Manzur, [The tongue of the Arabs] (in Arabic), DAR SADER, P. O. B. 10, Beirut, Lebanon, 1994.

[12] M. Zeki, O.O. Khalifa, A.W. Naji, “Development of an arabic text-to-speech system,” in: International Conference on Computer and Communication Engineering (ICCCE), 2010.

[13] I. A. Salim, [The syllabic structure in Arabic language] (in Arabic), Magazine of the Jordan Academy of Arabic 33, 1987, pp. 45–63.

[14] M. Alghamdi, A. H. Alhamid, M. M. Aldasuqi, “Database of Arabic Sounds: Sentences,” Technical Report, King Abdulaziz City of Science and Technology, Saudi Arabia, 2003. (In Arabic).

[15] [Holy Qur’an] (in Arabic: “القُرأنُ الكَرِيم”). [On-line]. Available: http://www.holyquran.net/quran/index.html [October 13, 2016].

[16] M. al-Bukhari. [Sahih al-Bukhari] (in Arabic: “صَحِيحُ البُخَارِي”). [On-line]. Available: http://shamela.ws/browse.php/book-1681 [October 13, 2016].

[17] S. Razi. [Nahj al-Balagha] (in Arabic: “نَهجُ البَلَاغَة”). [On-line]. Available: http://ia600306.us.archive.org/7/items/98472389432/nhj-blagh-ali.pdf [October 13, 2016].

[18] [The Mecca list of common vocabulary] (in Arabic: “قَائِمَةُ مَكَّةَ لِلمُفرَدَاتِ الشَّائِعَة”). [On-line]. Available: http://daleel-ar.com/2016/09/08/قائمة-مكة-للمفردات-الشائعة/ [October 13, 2016].

[19] K. Bobzin. [Arabic Basic Course] (in German: “Arabisch Grundkurs”). Wiesbaden, Germany: Harrassowitz Verlag, 2009.