Int J Speech Technol
DOI 10.1007/s10772-017-9440-2

The impact of phonological rules on Arabic speech recognition

Fawaz S. Al-Anzi[1] · **Dia AbuZeina**[1]

Received: 29 April 2017 / Accepted: 15 July 2017

Abstract The pronunciation variation is a well-known phenomenon that has been widely investigated for automatic speech recognition (ASR). The knowledge-based phonological rules are generally used to capture the accurate phonetic realization in order to minimize the mismatch between the ASR dictionary and the actual phonetic representation of the speech signal. For the Arabic ASR, there are a number of studies that employ these rules on Arabic ASR systems; however, little research has been devoted to measure the precise performance of each rule. In this paper, we aim at finding the exact effect of each rule as well as the rules that have no influence. We used the Carnegie Mellon University PocketSphinx speech recognizer with a new "in-house" modern standard Arabic speech corpus that contains 19 h for training and 3.7 h for testing. We evaluated the effect of three famous rules (Shadda, Tanween, and the solar letters). The experimental results do not show clear evidence that using phonological rules for ASR dictionary adaptation can enhance the performance for within-word pronunciation variation. The obtained results might be an indication to rethink or use other ASR performance aspects, such as cross-word pronunciation variation and the optimal phonemes set of the Arabic language.

Keywords Arabic · Speech recognition · Phonological rules · Sphinx

1 Introduction

Automatic speech recognition (ASR) is of particular interest in different fields, such as human computer interface (HCI) and the natural language processing (NLP). Recently, the Arabic large-vocabulary speaker-independent continuous speech recognition system has received significant attention in the NLP research community. However, Arabic ASR poses some challenges, such as the difficulty to obtain corpora for dialects that are spoken rather than written (i.e. there is no common standard for writing), difficulty in obtaining a large diacritized text as the Arabic allows writing without diacritics, and the enormous number of word forms due to the morphology richness of Arabic. In addition to the previous difficulties, the pronunciation variation phenomenon adds further challenges to ASR systems. That is, the continuous speech naturally has some acoustic variations that not accounted for in the pronunciation dictionary, which can lead to less than optimal performance. Due to the pronunciation variation problem, it is almost impossible to consider all possible variants in the pronunciation. No doubt that the mismatch between the acoustic features of the speech signal and the phonetic transcription in the ASR dictionary is a source of errors. In fact, it is extremely important that the phonemes of the pronunciation dictionary to adequately represent the actual contents of the training speech files. Pronunciation variations modeling is an active research area for robust ASR as well as the other related applications, such as text-to-speech systems to generate speech that is more natural.

One approach to tackle the pronunciation variations is through the language's phonological rules that consider the phonetic mismatch through ASR pronunciation dictionaries. For instance, (Ramsay et al. 2014) indicates that the performance of ASR is improved by shrinking the

✉ Fawaz S. Al-Anzi
FAWAZ.ALANZI@KU.EDU.KW

Dia AbuZeina
DIA.ABUZEINA@KU.EDU.KW

[1] Department of Computer Engineering, Kuwait University, Kuwait City, Kuwait

Published online: 24 July 2017

mismatch between the speech and the text used in training the acoustic model. Employing phonological rules for the ASR dictionary adaptation is classified as a knowledge based approach, however, data-driven is another option for the pronunciation variation. Hence, both approaches introduce some variants to generate phonetically rich dictionary pronunciation that might alleviate the acoustic changes on the performance. Modeling pronunciation variation includes two types, the within-word and the crossword pronunciation variation. In this work, we consider a knowledge based approach for within-word pronunciation variation. For comparison purposes and to evaluate the phonological rules, this work considers two testing cases: the baseline case and the dictionary adaptation case. The baseline case uses the phonemes set without any adaptation while the dictionary adaption case considers some rules. For each adaptation case, the ASR performance is evaluated to separately measure the impact of the corresponding rule.

In this paper, we employed the latest Carnegie Mellon University (CMU) PocketSphinx ASR engine (CMU Sphinx Downloads 2017) for exploring the effect of the Arabic phonological rule on the Arabic ASR performance. PocketSphinx includes the latest available releases as follows: sphinxbase—5prealpha, PocketSphinx—5prealpha, SphinxTrain—5prealpha. In the experiments, we used a new "in house" continuous speech corpus that contains 19 h for training and 3.7 h for testing. The speech is of broadcast news using the modern standard Arabic (MSA). The speech transcription is manually diacritized. This study also presents the intermediate steps for training and decoding, such as the proposed and used phonemes set, the pronunciation dictionary, the acoustic model, and the language model. We emphasize that this work is a preliminary step toward further research using the newly created corpus. This corpus has been fully supported by Kuwait University. The size of the corpus in this work is 22.7 h; however, we aim at increasing the size to about 30 h.

In the next section, we present the motivation of pronunciation variation for the Arabic ASR. In Sect. 3, we present the literature review followed by the phonemes set in Sect. 4. Section 5 presents the Arabic phonological rules followed by the baseline system in Sect. 6. Section 7 presents the proposed method and the experimental results in Sect. 8. We present diacritization in Sect. 9 and, finally, the conclusion and the future work are presented in Sect. 10.

2 Motivation

The acoustic properties of speech signals introduce some pronunciation variations, which is the major source of errors in ASR. Hence, employing phonological rules in ASR might enhance the supposed match between the transcription of the speech files and the actual acoustic features in the training process. In the case of training, without considering the phonological rules, many of the phonetic segments might lose suitable representation in the acoustic model. The differences between the actual speech signal and the phonetic spelling of the ASR dictionary leads to out-of-vocabulary word forms and, therefore, reduces the performance. The variation comes into the form of insertions, deletions, or substitutions of phoneme(s) beyond their listed forms in the ASR dictionary. (Benzeghiba and De Mori 2007) lists the major sources of errors in ASR, which include foreign and regional accents, speaker physiology, speaking style and spontaneous speech, rate of speech, children's speech, emotional state, and more. In order to handle the phonetic mismatch cases, some variants are generally added to the ASR dictionary (that is also called the lexical adaptation). For the Arabic ASR, little research has been devoted to find the exact contribution of each phonological rule on the overall performance. The motivation of this work is to explore the performance using some of the well-known rules. Based on our best knowledge, this is the first attempt to explore the effect of these rules using a continuous speech corpus. In fact, the Arabic ASR research is in need of exhaustive practical studies to define the most influential phonological rules. The precise evaluation of the most influential rules might lead to generating a phonetic transcription that is a reasonably approximation to reality. In addition, this study aims at finding the pronunciation rules that have no effect or even degrade the performance, if any.

3 Literature review

The literature shows that employing phonetically rich dictionaries will perform better than standard dictionaries that have no variants. For instance, (Fosler-Lussier et al. 1999) showed that the mismatch between the phonetics recognized and the word's phonetic transcription in the dictionary increases word error rate (WER) and degrades performance. (Fosler-Lussier et al. 1999) showed that the ASR performance will be highly improved if there is a closer match between the phonetic sequence recognized by the decoder and the phonetic transcription in the dictionary. Phonological rules have been utilized in ASR systems for different languages. For instance, (Tajchman et al. 1995) and (Finke and Waibel 1997) used a set of US English rules to generate pronunciation variants. (Wester 2003) and (Kessens et al. 1999) used a set of Dutch phonological rules to model pronunciation variations. (Kyong-Nim and Minhwa 2007) and (Jeon et al. 1998) used a set of Korean phonological rules to generate pronunciation variants. (Liu and Fung 2003) applied phonological rules to produce variants for Cantonese accented Mandarin speech. The knowledge-based approach was also implemented by (Seman and Jusoff 2008) for spontaneous Standard Malay.

For the Arabic language, (Ali et al. 2009) developed a software tool to generate pronunciation dictionaries for Arabic texts using Arabic pronunciation rules. This tool was later used in other works, such as (AbuZeina et al. 2011, 2012). However, the tool that was developed by (Ali et al. 2009) demonstrated the performance of the overall performance without the precise evaluation of each rule. (Alghamdi et al. 2007) demonstrates a phonetically rich ASR dictionary for a news transcription system for MSA. (Ramsay et al. 2014) presents a comprehensive system for generating a phonetic transcription based on a set of (language-dependent) pronunciation rules that convert the fully Arabic text into the actual sounds. The experimental results in (Abushariah et al. 2012) show that the non-diacritized case slightly outperforms the diacritized text case for a phonetically rich and balanced Arabic speech corpus. The research in (Vergyri et al. 2008) found that the diacritized text improved the acoustic model more than undiacritized orthography. Most of the previous works were performed using relatively small corpora; however, we used a larger corpus to explore the effect of phonological rules on Arabic ASR. (Masmoudi et al. 2014) employed a set of pronunciation rules (80 rules) for creating a phonetic dictionary for the Tunisian Arabic. (Biadsy et al. 2009) shows that using linguistically motivated pronunciation rules can significantly improves the ASR performance. (Al-Haj et al. 2009) demonstrated the knowledge-based approach to add variants to dictionary. They worked on the Iraqi-Arabic speech and focused on short vowels. The literature shows many studies have discussed the phonological rules, however, no study explores the impact of these rules separately.

4 The phonemes set

The phoneme is the basic unit of speech that represents a distinctive sound of the language's phonology. Hence, a change of a particular phoneme in a word makes a change in the meaning of the word. Phonemes play a vital role in the performance of ASR and text to speech systems. In this work, we propose a phoneme set that is used to evaluate the recognition performance of the prepared corpus. The pronunciation dictionary is prepared using the proposed phonemes set by a mapping process between the Arabic letters (the language's vowels and consonants) and their corresponding phonemes. However, in some cases, morphologically driven rules are used for a phonetic rich dictionary. In addition, some pronunciation exceptions might be manually processed for better acoustic representation. (Ali et al. 2009) and (Ramsay et al. 2014) elaborate on Arabic phonemes and the pronunciation rules.

In general, creating a dictionary of a particular language requires linguistic experts and a deep knowledge of the language sounds. However, the choice of the phoneme in the phonemes set is not straightforward as it has some constraints. For instances, it should represent the relevant information of the language sounds, it should consider the surrounding context between the letters, and it should carefully estimate the starting and the ending of the letters. No doubt, the phonemes that are used to represent the training

Table 1 The Arabic letters and the phonemes set

#	Letter	Phoneme	#	Letter	Phoneme	#	Letter	Phoneme
1	ء	E	17	ر	R	33	ه	H
2	آ	AA	18	ز	Z	34	و	W
3	أ	O	19	س	S	35	ى	AY
4	ؤ	EW	20	ش	SH	36	ي	Y
5	إ	I	21	ص	SS	37	◌ً	N
6	ئ	EY	22	ض	DD	38	◌ٌ	N
7	ا	A	23	ط	TT	39	◌ٍ	N
8	ب	B	24	ظ	ZZ	40	◌َ	AU
9	ة	P	25	ع	AE	41	◌ُ	AW
10	ت	T	26	غ	GH	42	◌ِ	AI
11	ث	TH	27	ف	F	43	◌ّ	~
12	ج	J	28	ق	Q	44	ـَا	AUA
13	ح	HH	29	ك	K	45	ـُو	AWW
14	خ	KH	30	ل	L	46	ـِي	AIY
15	د	D	31	م	M			
16	ذ	DH	32	ن	N			

words characterize the quality of the acoustic models and, therefore, the overall performance. Table 1 shows the phonemes set used in this work. It contains 46 phonemes. In addition to the Arabic letters, the table includes the short vowels that are Fatha (◌َ), Damma (◌ُ), and Kasra (◌ِ). As shown in the table, the Shadda (◌ّ) is represented using the symbol (~). We also used three phonemes to represent the Fatha that proceeds Alif (ا) ➔ ـَا as a single phoneme that is (AUA), the Damma that proceeds Waw (و) ➔ ـُو as a single phoneme that is (AWW), and the Kasra that proceeds Ya (ي) ➔ ـِي as (AIY). The reason for handling these cases as a single phoneme is that the pronunciation of the short vowels is different when it proceeds the long vowels. Hence, it would be correctly transcribed as single phonemes. For instance, "AW", "W" would be short vowel / AW/, consonant /W/, which is different in pronunciation from long /AWW/ as a single vowel, and likewise for the others. Consider the English name of the country "Kuwait". That would be correctly transcribed as /K AW W Y T/, because the /AW/ and /W/ are separate phonemes. But that's different from when the Arabic character "◌ُو" is used as the long vowel /AWW/ such as in the word "مَشرُوعَات" which means "projects". Hence, AUA, AWW and AIY has a better representation of the actual sounds that reflects the actual pronunciation. In this work, the transliteration of Arabic will be presented using the phonemes that are shown in Table 1. Table 1 has no symbol for Sukon (◌ْ) that does not correspond to any sound.

The selection of a phoneme symbol is an optional and it does not matter what phoneme symbol is used for an individual letter. For instance, (Ali et al. 2009) used (UW) for (ـُو) while we used (AWW). In the training stage, each phoneme is modelled using a sequence of a hidden Markov model (HMM) that is stated for computing the acoustic model. In the decoding stage, the phoneme is initially recognized and then used to find the most likely spoken words based on the best-matched phonemes between the speech file in question (the observations) and the trained HMMs of the acoustic model.

5 The Arabic phonological rules

In this work, we employ knowledge based phonological rules to model the pronunciation variations in an Arabic ASR for MSA. That is, a set of rules (defined by the experience of language experts) are used to adapt the phonetic dictionary in order to account for some variations that naturally occur in the Arabic pronunciation. (Elshafei 1991) is a good reference for the Arabic sounds. The essence of this work is to replace the standard phonetic representation to the expected actual pronunciations to, hopefully, perform better in the training and the decoding process. The rules convert the phonetic transcription in the dictionary to a "better" phonetic form that is close to the actual sounds based on the neighboring phonemes. Hence, the phonological rules could predict the variation within a word in order to control its representation in the dictionary. The rules introduced in this study include Shadda (الشدة), Nunation or Tanween (التنوين), and Assimilation (الادغام) using the sun letters also called solar letters (الحروف الشمسية).

To clarify the pronunciation differences due to the phonological rules, Table 2 shows the used rules along with examples. The Shadda (◌ّ) rule is a double or repeat of the previous consonant (also called the gemination mark). Nunation also called Tanween is a doubling of short vowels that includes (◌ُ: u, ◌َ: a, ◌ِ: i). Hence, Tanween includes any case of Dammatan (two consecutive short Damma), Fathatan (two consecutive short Fatha), or Kasratan (two consecutive short Kasra). Each Tanween is symbolized as (ـٌ ، ـً ، ـٍ). Assimilation is a merging of the sounds of two consecutive consonants (it could be within a single word or between two separated words) to produce a single geminated sound, so that the two sounds become alike or even identical. In this work, we used assimilation using the solar letters: {ت:T , ث:TH , د:D , ذ:DH , ر:R , ز:Z , س:S , ش:SH , ص:SS , ض:DD , ط:TT , ظ:ZZ , ل:L , ن:N } . The L "ل" that proceeds any of the solar consonants is assimilated with the consonant.

6 The baseline system

The goal of preparing the baseline system is to compare the performance when employing the phonological rules. Creating a continuous speech corpus was the first step in this research. We got the raw MSA speech files form (Al-Sabah TV 2007) in Kuwait. The speech contents belong to broadcast news. We performed the preprocessing step that includes segmenting the long speech files into short segments of 30–60 s. The produced segmented speech files cover different news stories and it sums up to 22.7 h of 29 speakers (19 male speakers and the rest are for female speakers). The speech files were sampled at 16 KHz mono. A silence of 0.1 s was used at the beginning and at the end of each speech file. We collected 2160 speech files that were transcribed and manually diacritized. The speech files were divided into two parts: the training set that contained 1802 speech files (19 h) and the testing set that contained 358 speech files (3.7 h). We emphasize that creating a continuous speech corpus is a time-consuming task.

The training stage of an ASR system consists of building an acoustic model that is a major component of ASR engines. Acoustic models statistically represent the relationships between the speech signals and the language phonemes. It has been long observed that the HMM based

Table 2 The popular Arabic phonological rules

No.	Rule	Examples	Phonetic transcription	Actual pronunciation	Meaning
1	Shadda	التّنمِيَة	A L T ~ AU N M AI Y AU P	التَتنمِيَة	Development
		التّنسِيق	A L T ~ AU N AU S AI Y Q	التَتَنسِيق	Formatting
		مُشَرِّف	M AW SH AU R ~ AI F	مُشَررِف	Honorable
2	Tanween	مُعضِلَةً	M AW AE DD AI L AU P WW	مُعضِلَتن	Problem
		مُعتبِرةً	M AW AE AU T B AI R P UU	مُعَتبِرتن	Consider
		مُفعَمَةٍ	M AW F AE AU M AU P II	مُفعَمَتن	Replete
3	Assimilation (solar letters)	التابِع	A L T A B AI AE AI	اتابِع	Dependent
		الثراءِ	A L TH AU R A E AI	اثراءِ	Get rich
		الداعِمَة	A L D A AE AI M AU P	اداعِمَة	Support
		الذَهَبِ	A L DH AU H AU B AI	اذَهَبِ	Gold
		الرَّئيسِ	A L R ~ AU EY Y S AI	ارَّئيسِ	President
		الزّجَاج	A L Z ~ AW J AU A J AI	ازّجَاج	Glass
		السّاحِلِ	A L S ~ A HH AI L AI	اسّاحِلِ	Coast
		الشَبَابُ	A L SH AU B AU A B AW	اشَبَابُ	Young
		الصَّادِر	A L SS ~ AU A D AI R	اصَّادِر	Issued
		الضَّبط	A L DD ~ AU B TT	اضَّبط	Settings
		الطَاقَة	A L TT ~ AU A Q AU P	اطَاقَة	Energy
		الظَن	A L ZZ ~ AU N	اظَن	Suspicion
		اللّجنَة	A L L AU ~ J N AU P	الّجنَة	Committee
		النَائِب	A L N AU A EY AI B	انَائِب	Deputy

acoustic models have been successfully implemented in the state of the art speech recognizers. CMU Sphinx speech engines support three types for HMM based acoustic modeling. For instance, the CMU Sphinx configuration file "Sphinx_train.cfg" has the commands to enable or disable the desired acoustic model. The types of acoustic models include the traditional fully continuous, the semi-continuous, and the phonetic tied-mixture (PTM) models. Despite the common implementation of fully continuous and semi-continuous in the Arabic ASR, however, PTM is a recent method that is compromised between important factors, such as speed and performance. It is also characterized by fast decoding as well as its ability to handle large amounts of speech collections. In this work, we used the PTM based acoustic models.

The pronunciation dictionary was generated using a Python based program based on the proposed phonemes set. The total number of unique words in the training set is 37,158. The corpus vocabulary and the size of the speech corpus determines some training parameters, such as the number of Senones (tied-state) and the number of Gaussians. Table 3 shows the approximation number of Senones and the Gaussian densities according to the vocabulary and the size of some English speech corpora (Training Acoustic Model for CMUSphinx 2017). For the language model, we used the CMU language toolkit (Building Language Model 2017) to calculate the statistical N-grams (i.e. 1, 2, and 3-g) based on the entire corpus transcription.

In addition to the previous steps, the SphinxTrain performs some internal tasks, such as computing features from audio files, training context independent models, training context dependent models, build trees, prune trees, lattice generation, lattice pruning, and finally decoding using the trained models. Once having the trained acoustic model, the PocketSphinx is used for decoding by utilizing other components, such as the pronunciation dictionary and the

Table 3 The approximate number of senones and Gaussian densities

Vocabulary	Hours	Senones	Densities	Example
20	5	200	8	Tidigits digits recognition
100	20	2000	8	RM1 command and control
5000	30	4000	16	WSJ1 5k small dictation
20,000	80	4000	32	WSJ1 20k big dictation
60,000	200	6000	16	HUB4 broadcast news
60,000	2000	12,000	64	Fisher rich telephone transcription

language model. In ASR, the training phase is time-consuming. Hence, we considered speeding up the execution time using an option in the PocketSphinx. We used the configuration file that is called "sphinx_train.cfg". This file has an option for the multiprocessing mode. The two options that can be used for reducing the training and the decoding time are as follows. $CFG_NPART = 10 ➔ the number of parts to run forward–backward estimation; and $DEC_CFG_NPART = 10 ➔ how many pieces to split decoding. The number ten is specified by the user according to the desired factor to reduce the execution time. The default value of these two parameters is one. This option is helpful since it clearly reduces the execution time by utilizing a number of processors in multicore machines.

7 The proposed method

The proposed method includes the dictionary adaptation (also called lexicon adaptation) process to change the phonetic transcription in the pronunciation dictionary according to the phonological rule effect. We have three cases that are included: Shadda, Tanween, and Assimilation. For the Shadda rule, we investigated two cases. The first case is to discard the Shadda. For instance, "T ~" becomes "T". The second case includes a replacement of the Shadda (~) to the proceeding consonant. For instance, "T ~" becomes "T T". Table 4 shows examples of the replacement process for the two cases. Of course, the replacement will occur for the dictionary's words that have Shadda. When implementing this rule, the phonemes set is also adapted to remove the Shadda phoneme (or the Shadda symbol) from the phonemes list since it is not used anymore. No change is performed on the language model. After the dictionary adaptation, the acoustic model is trained to generate the modified acoustic model.

The same process is repeated for the Tanween rule. For all the dictionary's entries that include any of the symbol {(WW :◌ٌ),(UU :◌ً),(II :◌ٍ)} they will be replaced to N. Hence, the Tanween rule is appended as N instead of the Tanween symbols. Table 5 shows some examples.

For the solar letters, the transformation includes an assimilation process of the phoneme (L) with the following solar consonant. (Akesson 2010) phonetically explained that the letter (L) and the solar letters have a close articulation area and they all originate from between the teeth to the lower part of the palate. Table 6 shows the dictionary adaptation for this rule.

Table 4 The Shadda rule transformation process

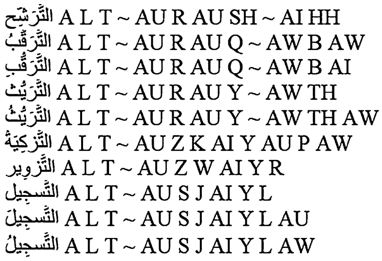
The Shadda rule	
Case 1: discarding Shadda	Case 2: duplicate the proceeding consonant
Before dictionary adaptation	Before dictionary adaptation
… أَجَّلَ O AU J ~ AU L AU …	… أَجَّلَ O AU J ~ AU L AU …
After dictionary adaptation	After dictionary adaptation
… أَجَّلَ O AU J AU L AU …	… أَجَّلَ O AU J J AU L AU …

Table 5 The Tanween rule transformation process

The Tanween rule		
Case 1: replace (◌ٌ) by (N)	Case 2: replace (◌ً) by (N)	Case 3: replace (◌ٍ) by (N)
Before dictionary adaptation	Before dictionary adaptation	Before dictionary adaptation
… أَعدَادٌ O AU AE D AU A D WW …	… أَعضَاءً O AU AE DD AU A E UU …	… أَهدَافٍ O AU H D AU A F II …
After dictionary adaptation	After dictionary adaptation	After dictionary adaptation
… أَعدَادٌ O AU AE D AU A D N …	… أَعضَاءً O AU AE DD AU A E N …	… أَهدَافٍ O AU H D AU A F N …

Table 6 The Solar letters transformation process

The Solar letters rule {ت:T , ث:TH , د:D , ذ:DH , ر:R , ز:Z , س:S , ش:SH , ص:SS , ض:DD , ط:TT , ظ:ZZ , ل:L , ن:N}		
Case 1: replace (التـ) by (اتـ)	Case 2 … Case 13 (for all Solar letters)	Case 14: replace (النـ) by (انـ)
Before dictionary adaptation	Before dictionary adaptation	Before dictionary adaptation
… التابِع A L T A B AI AE AI …	(ث)➔ الثراء, (د)➔ الدعم, (ذ)➔ الذهب, (ر)➔ الرسالة, (س)➔ السودان, (ش)➔ الشرق, (ص)➔ الصحافة, (ض)➔ الضخمة, (ط)➔ الطبي, (ظ)➔ الظروف, (ل)➔ اللائحة	… النَائِب A L N AU A EY AI B …
After dictionary adaptation	After dictionary adaptation	After dictionary adaptation
… التابِع A T A B AI AE AI …	(ث)➔ اثراء, (د)➔ ادعم, (ذ)➔ اذهب, (ر)➔ ارسالة, (س)➔ اسودان, (ش)➔ اشرق, (ص)➔ اصحافة, (ض)➔ اضخمة, (ط)➔ اطبي, (ظ)➔ اظروف, (ل)➔ الائحة	… النَائِب A N AU A EY AI B …



التَّرَشّح A L T ~ AU R AU SH ~ AI HH
التَّرَقّبُ A L T ~ AU R AU Q ~ AW B AW
التَّرَقّبِ A L T ~ AU R AU Q ~ AW B AI
التَّرَيُّث A L T ~ AU R AU Y ~ AW TH
التَّرَيُّثُ A L T ~ AU R AU Y ~ AW TH AW
التَّزكِيَةُ A L T ~ AU Z K AI Y AU P AW
التَّزوِير A L T ~ AU Z W AI Y R
التَّسجِيل A L T ~ AU S J AI Y L
التَّسجِيلَ A L T ~ AU S J AI Y L AU
التَّسجِيلُ A L T ~ AU S J AI Y L AW

Fig. 1 Some entries of the baseline dictionary

8 Experimental results

This section presents the experimental results based on the introduced MSA speech corpus. In this work, we used three emitting states of HMMs that corresponds to the subphones at the beginning, middle, and end of the phones. The acoustic models were calculated using context-dependent HMM triphones. Our acoustic models are all trained using the SphinxTrain for the phonetic tied-mixture (PTM) PocketSphinx. The performance is measured based on different parameters, such as the number of Senones and the number of Gaussian densities. Word Error Rate (WER) was used to evaluate the ASR performance in investigating the different cases. We initially evaluated the performance using the phonemes set presented in Table 1. Regarding the phonemes set, we evaluated the performance for two cases (43 phonemes and 46 phonemes). Figure 1 shows some entries of the baseline dictionary.

We initially conducted the experiments without employing the phonemes (ـَا:AUA, ـُو:AWW, ـِي:AIY). In fact, we wanted to measure the impact of these phonemes on the overall performance. Then, for the best performing case, we repeated an experiment with employing the phonemes (ـَا:AUA, ـُو:AWW, ـِي:AIY) as shown in Table 7. Hence, there is a slight performance difference when combining the short vowels with the long vowels. It seems that the difference (0.5%) is small but not significant. For investigating the phonological rules, we used the baseline dictionary that has no (ـَا:AUA, ـُو:AWW, ـِي:AIY) phonemes. It is worthy to mention that the Phonemes set also includes one more phoneme, which is SIL to handle the silent cases at the beginning and the end of speech files.

Table 7 The baseline system performance

Experiment	Densities	Senones	WER (%)	Accuracy (%)
43 Phonemes (without AUA, AWW, and AIY)				
1	64	500	31.2	68.8
2	128	500	31.2	68.8
3	256	500	31.2	68.8
4	64	1000	31.0	69.0
5	128	1000	30.9	69.1
6	256	1000	31.3	68.7
7	64	2000	31.1	68.9
8	128	2000	31.1	68.9
9	256	2000	31.5	68.5
46 Phonemes (with AUA, AWW, and AIY)				
10	128	1000	30.4	69.6

Table 8 Rough WERs for a number of English corpora

Speech collection	Vocabulary	WER %
TI Digits	11 (zero-nine, oh)	0.5
Wall Street Journal read speech	5000	3
Wall Street Journal read speech	20,000	3
Broadcast News	64,000+	10
Conversational telephone speech	64,000+	20

This relatively low accuracy is reasonable since we used a small size corpus. Ideally, ASR requires 200–300 h speech corpus. The language models also require huge textual data (gigabytes of text) for reasonable performance. It is reported in the Training Acoustic Model for CMUSphinx (2017) that the WER for 10-h task should be around 10%. For a large task, it could be around 30%. Table 8 shows the WER for some ASR systems using different English speech corpora (Jurafsky and Martin 2009).

One more reason for the obtained relatively low accuracy is that the used corpus has no filler dictionary. The filler dictionary generally contains noise and inhalation speech that are appropriately handled during the training phase. The fillers require indicating the noises and inhalations in the transcription of the speech files, which is an extremely difficult task for our corpus. The output of this work is demonstrated in Table 9. Based on the obtained results, no clear evidence that employed phonological rules is of significant performance enhancement in the case of a within-word pronunciation variation model. We emphasize that this work is based on replacing the standard phonetic transcription in the baseline dictionary by the proposed phonetic transcription, as some well-known method is based on adding the variants while, at the same time, keeping the standard phonetic transcription.

9 The effect of diacritization

Diacritization is the process of marking the letters using optional orthographic symbols that are called diacritics (i.e. the short vowels). The Arabic formal text is generally written without diacritics, which produces different pronunciation forms. That is, the Arabic writing system allows discarding short vowels and, hence, forcing the reader to use the prior knowledge and the words context to infer the missing diacritics. For the Arabic ASR, the problem of short vowels is that they are generally pronounced, but almost never written, which adds more challenges to the learning process. The missing of short vowels may increase the ambiguity in the acoustic model and, hence, produces less than optimal performance. The study in (Vergyri and Krichhoff 2004) indicates that the non-diacritized text leads to problems for both acoustic and language modeling and therefore may lead to a loss in recognition accuracy. Similarly, it is reported in (Kirchhoff et al. 2002) that the missing of short vowels leads to a significant increase in both the language model perplexity and the word error rate.

The importance of diacritization is that it enhances the supposed match between the phonetic transcription of the training textual files and the corresponding speech files. In fact, it is extremely important that the phonemes of the pronunciation dictionary adequately represent the actual training speech. In the case of training using non-diacritized text, many of phonetic segments will be lost because the short vowels are not there. Despite short vowels that help the reader to realize the meaning of a particular word, not using fully diacritized text might lead to ambiguity as the same word might have several meanings. For instance, the word “جنة: J N P” has three different meanings based on the short vowels (u:◌ُ,a:◌َ,i:◌ِ) on the first letter: (جُنة, جَنة, جِنة) (J U N P, J A N P, J I N P) so it can mean protection, paradise, and jinn, respectively. More on Arabic diacritization and some other related challenges are found in (Al-Anzi and AbuZeina 2015). On other hand, obtaining a sizable diacritized text for ASR and NLP applications is extremely difficult as well as a time-consuming task.

In this section, we present the performance using non-diacritized text. The experimental results show that the non-diacritized text system scored 81.2% while the diacritized text based system scored 69.1%. The dictionary size in case of non-diacritized is 23,481 unique words, however, the dictionary size of the base line is 37,158. Even the diacritized case has less accuracy due to the slight differences in diacritics; however, the non-diacritized case might be adequate and faultless for the Arabic native speakers. Regarding the execution time of both the training and the decoding stages, we found that the non-diacritized case required less execution time due to the reduced vocabulary.

10 Conclusion and future work

This paper presents an experimental ASR performance evaluation using a set of Arabic phonological rules. We investigated three well-known rules for within-word

Table 9 The Performance of the phonological rules

Experiment	The phonetic change	Densities	Senones	WER (%)	Accuracy (%)
The Shadda rule					
1	Remove ~	128	1000	30.4	69.6
2	Duplicate the proceeding	128	1000	31.2	68.8
The Tanween Rule					
3	Change (◌ٌ,◌ً,◌ٍ) to N	128	1000	31.0	69.0
The Solar letter rule					
4	Remove L before Solar letters	128	1000	30.7	69.3

pronunciation variations. We conducted the experiments using a new continuous speech corpus that contains about 22.7 h of news transcription. The corpus was manually diacritized. The experimental results reveal that employing the phonological rules does not clearly enhance the ASR performance. We investigated three rules that include Shadda, Tanween, and the solar letters. We emphasize that we replaced, and did not add, the phonetic transcription according to the phonological rules. Accordingly, the output of this work pushes to rethink the importance of phonological rules in ASR (i.e. for within-word pronunciation variations modeling). Hence, we recommend devoting ASR research for the cross-word pronunciation variations modeling as well as for finding the optimal phonemes set of the Arabic. In addition to the phonological rules, this paper presents an experimental evaluation of diacritized and non-diacritized based text. The experimental results show that the non-diacritized based system outperforms the diacritized based system even with a smaller vocabulary. However, the diacritized based system gives vowelized text output, which is not obtained by a non-diacritized based system.

Acknowledgements This work is supported by Kuwait University Research Administration Research Project Number EO06/12.

References

Abushariah, M. A.-A. M., et al. (2012). Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. *International Arab Journal of Information Technology, 9*(1), 84–93.

AbuZeina, D., et al. (2011). Toward enhanced Arabic speech recognition using part of speech tagging. *International Journal of Speech Technology, 14*(4), 419–426.

AbuZeina, D., et al. (2011). Cross-word Arabic pronunciation variation modeling for speech recognition. *International Journal of Speech Technology, 14*(3), 227–236.

AbuZeina, D., et al. (2012) Within-word pronunciation variation modeling for Arabic ASRs: A direct data-driven approach. *International Journal of Speech Technology, 15*(2), 65–75.

Akesson, J. (2010). *A study of the assimilation and substitution in Arabic*. Lund: Pallas Athena Distribution.

Al-Anzi, F. S., & AbuZeina, D. (2015). Stemming impact on Arabic text categorization performance: A survey. In *Proceedings of the 2015 5th international conference on information & communication technology and accessibility (ICTA)*, IEEE.

Alghamdi, M., Elshafei, M., & Al-Muhtaseb, H. (2007). Arabic broadcast news transcription system. *International Journal of Speech Technology, 10*(4), 183–195.

Al-Haj, H., Hsiao, R., Lane, I., Black, A., & Waibel, A. (2009) Pronunciation modeling for dialectal Arabic speech recognition, ASRU 2009: IEEE workshop, Italy.

Ali, M., Elshafei, M., Alghamdi, M., Almuhtaseb, H., & Alnajjar, A. (2009). Arabic phonetic dictionaries 236 for speech recognition. *Journal of Information Technology Research, 2*(4), 67–80.

Al-Sabah TV. (2017). http://www.alsabahpress.com/.

Benzeghiba, M., & De Mori, R. et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10–11), 763–786.

Biadsy, F., Habash, N., & Hirschberg, J. (2009). Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics.

Building Language Model. (2017). http://cmusphinx.sourceforge.net/wiki/tutoriallm.

CMU Sphinx Downloads. (2017). http://cmusphinx.sourceforge.net/wiki/download.

Elshafei, M.A. (1991) Toward an Arabic text-to-speech system. *The Arabian Journal for Science and Engineering, 16*(4B), 565–583.

Finke, M., & Waibel, A. (1997). Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proceedings of EuroSpeech-97* (pp. 2379–2382), Rhodes.

Fosler-Lussier, E., Greenberg, S., & Morgan, N. (1999) Incorporating contextual phonetics into automatic speech recognition. In *Proceedings of the international congress on phonetic sciences*, (pp 611–614).

Jeon, J., Cha, S., Chung, M., Park, J., & Hwang, K. (1998). Automatic generation of Korean pronunciation variants by multistage applications of phonological rules. In *ICSLP-1998* (paper 0675).

Jurafsky, D., Martin, J. (2009). *Speech and language processing*, 2nd edn. Hoboken: Pearson.

Kessens, J. M., Wester, M., et al. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication, 29*(2–4), 193–207.

Kirchhoff, K., et al. (2002) Novel approaches to Arabic speech recognition-final report from the JHU summer workshop 2002. Technical Reports, John-Hopkins University.

Kyong-Nim, L. & Minhwa, C. (2007). Morpheme-based modeling of pronunciation variation for large vocabulary continuous speech recognition in Korean, *IEICE Transactions on Information and Systems, 90*(7), 1063–1072.

Liu, Y., & Fung, P. (2003). Modeling partial pronunciation variations for spontaneous Mandarin speech recognition. *Computer Speech and Language, 17*, 357–379.

Masmoudi, A., et al. (2014) A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition. In *LREC*.

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.

Seman, N., & Jusoff, K. (2008). Acoustic pronunciation variations modeling for standard Malay speech recognition. *Computer and Information Science, 1*(4), 112.

Tajchman, G., Foster, E., Jurafsky, D. (1995) Building multiple pronunciation models for novel words using exploratory computational phonology. In *EUROSPEECH-1995* (pp. 2247–2250).

Training Acoustic Model for CMUSphinx. (2017). http://cmusphinx.sourceforge.net/wiki/tutorialam.

Vergyri, D., et al. (2008) Development of the SRI/nightingale Arabic ASR system. Interspeech.

Vergyri, D., & Kirchhoff, K. (2004). Automatic diacritization of Arabic for acoustic modeling in speech recognition. In *Proceedings of the workshop on computational approaches to Arabic script-based languages*, Association for Computational Linguistics.

Wester, M. (2003). Pronunciation modeling for ASR: Knowledge-based and data-derived methods. *Computer Speech & Language, 17*, 69–85.