Handling within-word and cross-word pronunciation variation for Arabic speech recognition (knowledge-based approach)

Ibrahim El-Henawy, Marwa Abo-Elazm*

Computer Science Department Faculty of Computers and Informatics Zagazig University Egypt
Emails: henawy2000@yahoo.com, marwa_abdella@yahoo.com

Abstract

Arabic is one of the phonetically complex languages, and the creation of an accurate speech recognition system is a challenging task. The phonetic dictionary is an essential component in an automatic speech recognition system (ASR). The pronunciation variations in Arabic are tangible and are investigated widely using a data-driven approach or knowledge-based approach. The phonological rules are used to get the pronunciation of each word accurately to reduce the mismatch between the actual phoneme representation of the spoken words and the ASR dictionary. Several studies in the Arabic ASR system are conducted using a different number of phonological rules. In this paper, we focus on those rules that handle within-word pronunciation variation and cross-word pronunciation variation. The experimental results indicate that handling within-word pronunciation variation using phonological rules doesn't enhance recognition performance, but using these rules to handle cross-word variation provides good performance.

Keywords: Speech Recognition Systems; Arabic Language; Phonetic Dictionary; pronunciation variations

1. Introduction

Automatic speech recognition can be defined as converting the speech signal into text. The quality of the system is measured by knowing how much the recognized text is close to the text recognized by a human. Speech recognition is taking a large interest in many fields such as natural language processing (NLP) and human-computer interaction (HCI). There are three types of Arabic language each has different characteristics [1]: Classical Arabic (CA), which is the formal and standard form of Arabic, is the Quran language, Modern Standard Arabic (MSA), used in TV and the news the "common language" used by speakers of different dialects, and Spoken Arabic (dialect), that differ from one country to other and have no organized writing form. Despite the importance of the Arabic language and the research effort, Arabic Automatic Speech Recognition (ASR) is unfortunately still insufficient. Several issues in the Arabic language need to be addressed to catch up with the progress of other languages [2]. Discretization is one of the obstacles faced by Arabic ASR systems, science not all text is discretized and this lead to a shortage in the training data needed by ASR systems. Discretization is essential for the Arabic ASR system that is integrated with other systems in which this system performs better using diacritics such as speech-to-speech systems[3]. The other problem is morphological complexity since Arabic has a large potential for word forms that increases the out-vocabulary rate. Also, pronunciation variations (within a word or crossword variation) lead to a mismatch between the spoken word and the text used in the ARS system modeling. Within-word variation causes alternate pronunciations of the same word. In contrast, a cross-word variation happens in a continuous speech in which a sequence of words forms a compound word that must be treated as one entity [4]. Modeling the pronunciation

DOI: https://doi.org/10.54216/JISIoT.010202

variation in any ASR system is a critical task. It helps to improve performance by reducing the mismatch between the speech and the text used in the acoustic model training [5][6].

Two main methods used in the previous literature in modeling the pronunciation -variation [7][8] Knowledge-based approach, which uses phonetic and linguistic knowledge to write phonological rules that handle variants in pronunciation. The data-driven approach uses a corpus from real speech to derive the variation in speech. The chosen approach depends on the type of variation you need to handle in your work and the purpose of handling these variations [6]. The pronunciation variation modeling should be considered on three levels: the pronunciation dictionary, the acoustic model, and the language model [9].

2. Arabic phoneme set

The phoneme is the small and basic unit of speech. It represents a distinct sound of the language's phonology. Any phoneme change 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 used a phoneme set that is used in [10] in addition to the proposed phoneme to generate the adapted dictionary to handle word variation. The Arabic language contains 28 consonants, 3 short vowels representing Fatha, Damma, and Kasra,3 long vowels that are the long version of the short vowels, and the pharyngealized allophone as illustrated in table 1.

Table 1: The phoneme set used in training

Number	Arabic phoneme	Romanized phoneme	Description	Number	Arabic phoneme	Romanized phoneme	Description
1	ـَ	/AE/	diacritical marks FATHA	24	ذ	/DH/	Arabic consonant ZAL
2	اَ	/AE:/	long vowel of AE	25	ر	/R/	Arabic consonant RA
3	ـَ	/AA/	the pharyngeal allophone of /AE/	26	ز	/Z/	Arabic consonant ZA
4	ـَ	/AA:/	Long version of AA	27	س	/S/	Arabic consonant SEEN
5	ـَ	/AH/	Emphatic Version of /AE/	28	ش	/SH/	Arabic consonant SHEEN
6	ـَ	/AH:/	The long version of AH	29	ص	/SS/	Arabic consonant SAD
7	ـٌ	/UH/	Diacritical marks DHAMMA	30	ض	/DD/	Arabic consonant DAD
8	وُ	/UW/	Long vowel UH	31	ط	TT	Arabic consonant TA
9	ـُ	/UX/	the pharyngealized allophone of /UH/	32	ظ	DH2	Arabic consonant THA
10	ـِ*	/IH/	diacritical marks KASSRA	33	ع	AI	Arabic consonant AIN
11	يِ	/IY/	Long vowel of IY	34	غ	GH	Arabic consonant GHAIN
12	ـِ*	/IX/	the pharyngeal allophone of /IH/	35	ف	F	Arabic consonant FA
13	ـِ*	/IX:/	The long version of IX	36	ق	Q	Arabic consonant QAF
14	وَ	/AW/	A Diphthong of both /AE/	37	ك	K	Arabic consonant KAF

DOI: https://doi.org/10.54216/JISIoT.010202

			and /UH/				
15	ىَ	/AY/	A Diphthong of both /AE/ and /IH/	38	ل	L	Arabic consonant LAM
16	ء	/E/	Hamza	39	م	M	Arabic consonant MEM
17	ب	/B/	Arabic consonant BA	40	ن	N	Arabic consonant NON
18	ت	/T/	Arabic consonant TA	٤١	ن	NN	Arabic consonant NON
19	ث	/TH/	Arabic consonant THA	٤٢	ن	NK	Arabic consonant NON
20	ج	/JH/	Arabic consonant GEEM	٤٣	ن	NF	Arabic consonant NON
21	ح	/HH/	Arabic consonant HA	4٤	ه	H	Arabic consonant HA
22	خ	/KH/	Arabic consonant KHA	4٥	و	W	Arabic Semi-vowel WAW
23	د	/D/	Arabic consonant DAL	4٦	ى	Y	Arabic Semi-vowel YA

3. The Arabic phonological rules

The phonetic dictionary has a great impact on the accuracy of the ASR system, it contains the words available in the language and their pronunciation as phonemes or allophones exist in the acoustic model. the dictionary creation can be done manually by an expert but it's a hard task and take a big time, for example, English dictionary is built manually over many years because of large exceptions [10]. The pronunciation of the Arabic language follows specific rules especially when the text is fully discretized, so the creation of the phonetic dictionary can be done automatically following these rules [11-12]. After the dictionary generation, it can be adapted manually for exception words. a number of research issues for Arabic speech recognition such as the absence of short vowels in written text and the presence of compound words generated from the concatenation of conjunctions, prepositions, articles, and pronouns, as prefixes and suffixes to the word stem is discussed in[13]. An Arabic broadcast news transcription system is developed and its phonetic dictionary provides different pronunciation variations for words that may be pronounced differently[14]. A change to the standard phonetic rule to adapt the pronunciation variation for better training and decoding process is developed. A rule-based technique is developed to generate Arabic phonetic dictionaries for a large vocabulary speech recognition system [10]. They used classic Arabic pronunciation rules, MSA rules, and morphologically driven rules. Al-Haj et al. (2009) create a knowledge-based approach to handling short vowels for Iraqi-Arabic speech and a number of pronunciation variations in the phonetic dictionary. A set of 80 pronunciation rules is generated to create a phonetic dictionary for Tunisian Arabic [15].

4. The proposed method

Some letters have different pronunciations when followed by a special letter such as the letter DAL[d], THE[t], and DAD[dd] for example the letter DAL(د) when followed by a voweled TEH(ت) it is omitted also, the letter DAD (ض) when followed by a voweled TEH(ت) or TAH(ط) is omitted [2], but Ramsay et al. (2014) made modifications to this rules in which the letter DAL(د) is pronounced as [t] when followed by a voweled TEH(ت) also, letter TEH(ت) is pronounced as [d] when followed by letter DAL(د). Ali et.al (2009) indicate that the consonant noon(ن)is partially assimilated into meem(م) when followed by baa(ب) For example the word مِنبَر(M E N B A R) is pronounced [M E M B A R], but Ramsay et .al (2014)add another rule for the letter noon (ن) see figure 1.

DOI: https://doi.org/10.54216/JISIoT.010202

NOON:

.(?=BEH) -> M

.(?= FEH) -> ɱ

.(?= (QAF|KAF)) ->ŋ

.(?=PN) ->n̟

Figure 1: NOON assimilations rule

It indicates that this letter adopts the labiality of the consonant FEH [f] and is assimilated to [ɱ] every time it is followed by FEH [f] For example the word يَنفَذ (Y A N F A Z) is pronounced[Y A ɱ F A Z], also it adopts the verity of the sounds KAF [k] and QAF [q] when followed by one of them and is pronounced as [ŋ]. For example, the word بَنك (B A N K) is pronounced [B A n̟ K]. it is assimilated to the palatal nasal consonant [n̟] when followed by one of the sounds articulated from the postal veolar points and dental, both stops and fricatives(PN), namely: DAD(ض), TAH(ط), ZAH(ظ), ZAIN(ز), SEEN(س), SHEEN(ش), JEEM(ج), THEH(ث), THE(ت), THAL(ذ), DAL(د), ZAIN(ز). For example, the word مَنثُورَ (M A N TH O R A) is pronounced [M A n̟ TH O R A]. Table2 shows the phonetic dictionary for the two approaches in which point view 1 for the approach in [5]and point view 2 for the approach in [2]. The phoneme NN, NK, and NF are used to represent n̟, ŋ, and ɱ respectively.

Table 2: Arabic phonological rules

Rule	example	Point view1	Point view2
DAD asmiliation	أَفَضتُم	E AE F AE DD T UH M	E AE F AE T UH M
DAL asmiliation	عُدتُ	AI UH T T UH	AI UH T UH
Shadda	العِرقِيُّ	E L AI IH R Q IX Y Y UH	E L AI IH R Q IX Y UH
Noon Asmiliation	إِنسَانٍ	E IH NN S AE: N	E IH N S AE: N
	العَنكَبُوتِ	E L AI AE NK K AE B UW T IH	E L AI AE N K AE B UW T IH
	العُنفِ	E L AI UH NF F IH	E L AI UH N F IH

Ramsy et al. (2014) investigate the noon nasalization rule with another rule such as Shadda(الشدة)and tanween(التنوين) on small corpora 20 sentence , indicating that the total performance is enhanced, but Al–Anzi Fawaz et al (2017) investigate the Shadda(الشدة) and tanween(التنوين) separately indicating that employed phonological rules are of no significant performance enhancement

The proposed method generates the dictionary adaption using the phonological rules used in Ramsy et al. (2014) duplicate the letter with shadda for example بّ become b b and the assimilation of N TO NN ,NF or NK.

DOI: https://doi.org/10.54216/JISIoT.010202

5. Experiment result

This experiment is conducted using the Nawar Halabi dataset which is a continuous speaker-dependent speech corpus. The transcript of the dataset was collected from Aljazeera Learn (Al Jazeera, 2015), which is a language learning website that was chosen because it contained fully discretized text which makes it easier to phonetize. The training data consist of 1813 files about 3.8h. The test data consists of 100 files in about 18 minutes. Carnegie Mellon University (CMU) sphinx 3 ASR engine is used. The WAV files are resampled to match the engine requirements i.e. sampling rate is 16 kHz, 16 bits and Wave format is Mono wav. The engine uses 3 state Hidden Markove Model. All wav file is fully transcribed in full discretized text. A transcript file tells the trainer which unit sounds it should learn the parameters of, and at least their order in every speech signal in the training. It contains a series of words and non-speech sounds written according to their order in a speech signal, followed by a tag to join this order with the corresponding speech signal. The baseline system used in this experiment is the system provided by Ali et al. (2009). The acoustic model is trained using the sphinxtrain continuous dependent (CD) Hidden Markove Model. The performance of the model is affected by the number of densities, so the experiment is repeated for 8, 16, 32, and 64 densities, the smallest WER is 11.27% achieved using 16 densities see figure 2. Two experiments are conducted using the number of densities that has the small WER to verify the two rules: the nasalization of noon (ن) and Shadda (الشدة) rule. The two -experiments show that no enhancement is done with these rules: duplicating phoneme that proceeds shadda and this is the same result obtained by Al–Anzi Fawaz et al (2017), also noon assimilation doesn't enhance the performance of the ASR system see table 4.

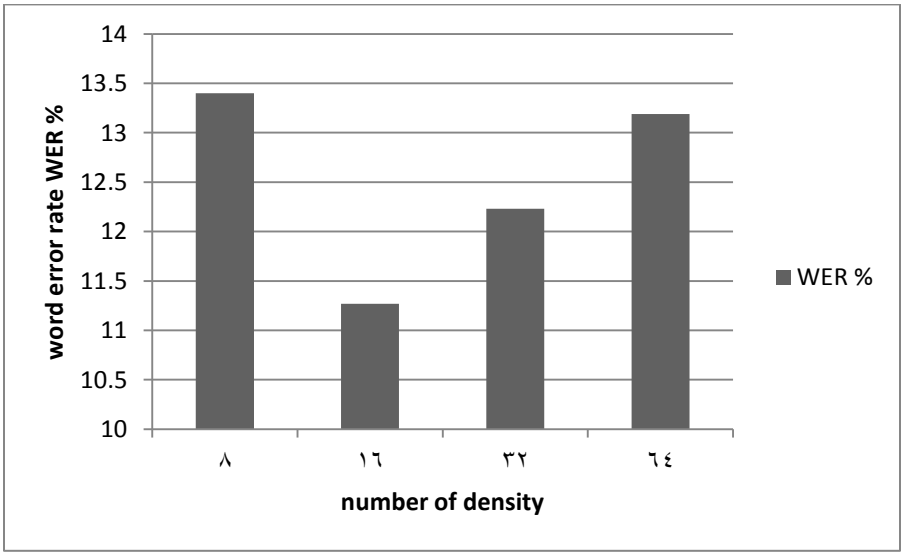

Figure 2: the word error rate (WER) for different numbers of density

6. Handling Cross-word pronunciation variations

Cross-word pronunciation variations change the phonetic spelling of words outside their listed forms in the phonetic dictionary, this leads to a number of Out-Of-Vocabulary (OOV) word forms [8]. The cross-word variation occurs at the intervals of words that are captured by the triphones of the acoustic model. It could also be realized as a change in pronunciation according to the last phoneme of a word and the first phoneme of the next word [16]. While cross-word variation modeling has been done in many Languages, little work in Arabic is done. Two well-known MSA phonological rules are applied, assimilation (Idgham) and changing (Iqlaab).

DOI: https://doi.org/10.54216/JISIoT.010202

There are 3 types of assimilation

- Noon Saakinah or Tanween

It is the merging between Noon Saakinah (ن) or Tanween(ً, ٍ ,ٌ)with one of the following letters (ي ر م ل و ن). For example the sentence وَمِنَ الْمُنْتَظَرِ أَيَّلْتَقِيَ فِي عَمَّانَ قِيَادَاتٍ دِينِيَّةً إِسْلَامِيَّةً وَمَسِيحِيَّةً becomes وَمِنَ الْمُنْتَظَرِ أَيَّلْتَقِيَ فِي عَمَّانَ قِيَادَاتٍ دِينِيَّةً إِسْلَامِيَّةً وَمَسِيحِيَّةً

- Assimilation of identical letters

An unvoweled consonant on the end of a word is merged with a same voweled consonant at the beginning of the next word to produce a new double consonant.For example the sentence وَبَلَغَتْ تَكْلِفَةُ الْمَبَارَّةِ مِائَةَ الْف جُنَّيهِ becomes وَبَلَغَتْكْلِفَةُ الْمَبَارَّةِ مِائَةَ الْف جُنَّيهِ

- Assimilation of two close pronunciation letters

Two successive different letters that are close in pronunciation as indicated in table 3 are merged and becoming one doubled letter, for example the sentence إضطرَّت طَائِرَةُ مُورَالِيس لِلْهُبُوطِ بِمَطَارِ فِيينَّا بَعْدَ رَفْضِ دُوَلٍ أُورُوبِيَّةٍ become إضطَرَّطَّائِرَةُ مُورَالِيس لِلْهُبُوطِ بِمَطَارِ فِيينَّا بَعْدَ رَفْضِ دُوَلٍ أُورُوبِيَّةٍ .

Table 3: Idgham of two close in pronunciation letters

Last letter Of first word Unvowelled	First letter of second word Vowelled	Connecting letter **Double**
TEH (ت)	DAL(د)	DAL(د)
THE(ت)	TAH(ط)	TAH(ط)
DAL(د)	THE(ت)	THE(ت)
BA(ب)	MEM(م)	MEM(م)
DH (ذ)	DH2(ظ)	DH2(ظ)
KAF(ك)	QAF(ق)	QAF(ق)
LAM(ل)	RA(ر)	RA(ر)
THA(ث)	THAL(ذ)	THAL(ذ)
TAH(ط)	THE(ت)	THE(ت)

The Iqlaab is a replacement of Noon Saakinah (ن) or Tanween (ً,ٍ,ٌ) that followed by voweled Baa (ب) with Meem Saakinah (م).for example the sentence وَنَصَحَتِ الدِّرَاسَةُ بِمُمَارَسَةِ الغِنَاءِ حَتَّى وَإِن لَم يَكُن بِشَكلٍ مُتقَنٍ becomes وَنَصَحَتِ الدِّرَاسَةُ بِمُمَارَسَةِ الغِنَاءِ حَتَّى وَإِن لَم يَكُمبِشَكلٍ مُتقَنٍ

Handling tanween is different not only removing tanween sign(ً,ٍ ,ٌ) from the letter but also adding a vowel (َ,ِ,ُ)on the letter according to the tanween letter, for example خُطُوطٍ مُلَائِمَةٍ become خُطُوطِمُلَائِمَةٍ and ارتِبَاطٌ رُوحِيٌّ become ارتِبَاطُرُّوحِيٌّ, also when handling tanween with fath (ً)there is an extra alif (ا) that need to be removed before the merging process, for example وِفقاً لِمَفهُومٍ become وِفقَلِّمَفهُومٍ. In this experiment, this modeling is handled in all ASR components the dictionary acoustic model and the language model. First phonetic adaption is done followed by adapting language and the adapted model is used in training the acoustic model to generate model parameters. The proposed model outperforms the baseline with an error rate 10.47% and the execution time is the same as the base system execution time as shown in table 4, where I: Word Insertion errors; D: word deletion errors; S: word substitution errors; WER: % word error rate, based on 1251word in 100 sentence test corpus.

DOI: https://doi.org/10.54216/JISIoT.010202

Table 4: the performance of the ASR system for different test cases

Test case		WER %	**I**	**D**	**S**	**Execution time in seconds**
Base system		11.27	**31**	**4**	**106**	**30.3**
Within word variation	Noon assimilation	11.43	**30**	**3**	**110**	**30.3**
	Shadda and noon assimilation	12.15	**40**	**2**	**110**	**40.4**
Crossword variation		10.47	**31**	**5**	**95**	**30.3**

7. Conclusion

Handling Arabic pronunciations variation influence Arabic ASR systems performance. Two types of variation exist which are within-word variations and cross-word variations. Handling within-word variation (Noon assimilation and shadda) using phonological rules (the knowledge-based approach) has no significant effect on system performance. On the other hand, better performance is achieved when handling cross-word variation by phonological rules. Accordingly, handling within-word variation using a data-driven approach need to be examined. Also, more phonological rules to handle other cross-word variations will be checked.

REFERENCES

[1] Elmahdy et al. used acoustic models trained with a large MSA news broadcast speech corpus to work as multilingual or multi-accent models to decode colloquial Arabic(2009).

[2] Al–Anzi Fawaz S, AbuZeina Dia, "The impact of phonological rules on Arabic speech recognition", International Journal of Speech Technology, vol. 20, no. 3, pp. 715-723, 2017.

[3] Abed, S., Alshayeji, M. and Sultan, S. 2019. Diacritics Effect on Arabic Speech Recognition. Arabian Journal for Science and Engineering. (2019).

[4] Abuzeina, D., Al-Khatib, W., Elshafei, M., Al-Muhtaseb, H., 2011. Cross-word Arabic pronunciation variation modeling for speech recognition.Int. J. Speech Technol. 14 (3), 227–236.

[5] Fosler-Lussier, E., Greenberg, S., Morgan, N., et al., 1999. Incorporating contextual phonetics into automatic speech recognition. Nucleus 48993(65.3), 62118.

[6] 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.

DOI: https://doi.org/10.54216/JISIoT.010202

[7] Amdal, I., Fosler-Lussier, E., 2003. Pronunciation variation modeling in automatic speech recognition. Telektronikk 99 (2), 70–82.

[8]Wester, M., Fosler-Lussier, E., 2000. A comparison of data-derived and knowledge-based modeling of pronunciation variation.

[9] Helmer, S. (2001). Pronunciation adaptation at the lexical level. In Proceedings ISCA ITRW workshop adaptation methods for speech recognition, Sophia Antipolis, France.

[10] Ali, M., Moustafa, E., Mansour, A., Husni, A., & Atef, A. (2009). Arabic phonetic dictionaries for speech recognition. Journal of Information Technology Research, 2(4), 67–80.

[11] Algamdi, M., Almuhtasib, H., & Elshafei, M. (2004).Arabic Phonological Rules. [King Saud University.].Journal of Computer Sciences and Information,16, 1–25.

[12] Elshafei-Ahmed, M. (1991). Toward an Arabic Text-to-Speech System. The Arabian Journal of Scienceand Engineering, 16(4B), 565–583.

[13] Billa et al. (2002). Arabic speech and test in tides on tap. In Proceedings of HLT.

[14] Alghamdi, M., Elshafei, M., & Almuhtasib, H. (2009). Arabic broadcast news transcription system. International Journal of Speech and Technology, 10, 183–195.

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

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

DOI: https://doi.org/10.54216/JISIoT.010202