

Generation of Arabic Phonetic Dictionaries for Speech Recognition

Mohamed Ali*, Moustafa Elshafei*, Mansour Al-Ghamdi**, Husni Al-Muhtaseb*, and Atef Al-Najjar*

**King Fahd University of Petroleum and Minerals, **King Abdulaziz City of Science and Technology*

mohamed-a@live.com, {elshafei, muhtaseb, alnajjar}@kfupm.edu.sa, mghamdi@kacst.edu.sa

Abstract

Phonetic dictionaries are essential components of large-vocabulary natural language speaker-independent speech recognition systems. This paper presents a rule-based technique to generate Arabic phonetic dictionaries for a large vocabulary speech recognition system. The system used classic Arabic pronunciation rules, common pronunciation rules of Modern Standard Arabic, as well as morphologically driven rules. The paper gives in detail an explanation of these rules as well as their formal mathematical presentation. The rules were used to generate a dictionary for a 5.4 hours corpus of broadcast news. The phonetic dictionary contains 23,841 definitions corresponding to about 14232 words. The generated dictionary was evaluated on an actual Arabic speech recognition system. The pronunciation rules and the phone set were validated by test cases. The Arabic speech recognition system achieves word error rate of %11.71 for fully diacritized transcription of about 1.1 hours of Arabic broadcast news.

1. Introduction

Automatic Speech Recognition (ASR) is a key technology for a variety of industrial and IT applications; it extends the reach of IT across people as well as applications. Automatic Speech Recognition (ASR) is gaining a growing role for a variety of applications, such as hands-free operation and control (as in cars and airplanes), automatic query answering, telephone communication with information systems, automatic dictation (speech-to-text transcription), government information systems, etc. In fact, speech communication with computers, PCs, and household appliances is envisioned to be the dominant human-machine interface in the near future. In spite of the tangible success of this technology for the English language and other languages, there are still many issues need to be addressed by researchers to reach the same level for the Arabic language.

One of the key components of the modern large vocabulary speech recognition systems is the

pronunciation or phonetic dictionary [1, 2, and 3]. This dictionary serves as an intermediary between the Acoustic Model and the Language Model in speech recognition systems. It contains the words available in the language and the pronunciation of each in terms of the phonemes or the allophones available in the acoustic model.

The Carnegie Mellon University (CMU) open source dictionary for North American English contains over 125,000 words and their transcriptions. The format of this dictionary is particularly useful for speech recognition and synthesis, as it has mappings from words to their pronunciations in a given phoneme set. The current phoneme set of the CMU dictionary contains 39 English phonemes [3]. Because of the large number of pronunciation exceptions in English, this dictionary was mainly built manually by experts over many years.

On the other hand, pronunciation of Arabic language follows specific rules and patterns if the provided text is fully diacritized. Many of these pronunciation rules can be found in [4] and [5].

The statistical approach for speech recognition is dominated by a powerful statistical technique called Hidden Markov Model (HMM) [1]. The HMM-based ASR technique has led to numerous applications requiring large vocabulary speaker-independent continuous speech recognition [1, 2].

The HMM-based technique essentially consists of recognizing speech by estimating the likelihood of each phoneme at contiguous, small frames of the speech signal [7, 8]. Words in the target vocabulary are modeled into a sequence of phonemes, and then a search procedure is used to find, amongst the words in the vocabulary list, the sequence that best matches the sequence of phonemes of the spoken word. Two notable successes in the academic community in developing high performance large vocabulary speaker independent speech recognition systems over the last two decades are the HTK tool kit, developed at Cambridge University [7]; and the Sphinx system developed at Carnegie Mellon University [8].

Development of an Arabic speech recognition is a multi-discipline effort, which requires integration of Arabic phonetic [9], Arabic speech processing techniques [10, 11], and Natural language processing

[12]. Development of an Arabic speech recognition system has recently been addressed by a number of researchers. Saroti et. al., 2007 [13] used Sphinx tools for Arabic speech recognition. They demonstrated the use of the tools for recognition of isolated Arabic digits. The data was recorded from 6 speakers. They achieved digits recognition accuracy of 86.66%. Hiyassat, 2007 [14], in his Ph.D. thesis developed a tool to generate Arabic pronunciation dictionaries. The generated dictionaries are based on a small MSA speech corpus consisting of digits or command and control vocabulary.

A workshop was held in 2002 at John Hopkins University [15] to define and address the challenges in developing a speech recognition system using Egyptian dialectic Arabic for telephone conversations. They proposed to use Romanization method for transcription of the speech corpus.

All the current approaches for large vocabulary Arabic speech recognition systems require in the first place an Arabic phonetic dictionary together with its management tools. In the following sections we provide in detail the rules to be used for automatic generation of such dictionary from the transcription of a given Arabic speech corpus. We also show how the method is tested and validated using an actual large vocabulary Arabic speech recognition system.

In Section 2, we provide a brief introduction to the phoneme/allophone set used in building the dictionary. Then in Section 3 we describe in detail the rules used in building the dictionary. Section 4 covers the Arabic speech corpus used in testing the rules. Finally, in Section 5, we present the testing and evaluation results of the proposed method.

2. Arabic Phoneme Set

Table 1 shows the listing of the phoneme set used in the training and the corresponding phoneme symbols. The table also shows illustrative examples of the vowel usage. The phoneme set chosen is based on the previous experience with Arabic text-to-Speech systems [4, 5, and 10], and the corresponding phoneme set which is successfully used in the English ASR [3].

Table 1. The complete phoneme set used in training

Phoneme	Letter	Phoneme	Letter
/AE/	ا (Fatha)	/KH/	خ (Khah)
/AE:/	آ (Damma)	/D/	د (Dal)
/AA/	أ (Kasra)	/DH/	ذ (Thal)
/AH/	هـ (Hamza)	/R/	ر (Reh)
/UH/	و (Damma)	/Z/	ز (Zain)
/UW/	وو (Damma)	/S/	س (Seen)
/UX/	ؤ (Kasra)	/SH/	ش (Sheen)
/IH/	ي (Kasra)	/SS/	ص (Sad)
/IY/	ي (Kasra)	/DD/	ض (Dad)
/IX/	ي (Kasra)	/TT/	ط (Tah)
/AW/	او (Damma)	/DH2/	ظ (Thah)

/AY/	آ (Damma)	/AI/	ع (Ain)
/UN/	ن (Noon)	/GH/	غ (Ghain)
/AN/	ن (Noon)	/F/	ف (Feh)
/IN/	ن (Noon)	/V/	ف (Feh)
/E/	ء (Hamza)	/Q/	ق (Qaf)
/B/	ب (Beh)	/K/	ك (Kaf)
/T/	ت (Teh)	/L/	ل (Lam)
/TH/	ث (Theh)	/M/	م (Meem)
/JH/	ج (Jeem)	/N/	ن (Noon)
/G/	ج (Jeem)	/H/	هـ (Heh)
/ZH/	ج (Jeem)	/W/	و (Waw)
/HH/	ح (Hah)	/Y/	ي (Yeh)

The regular Arabic short vowels are /AE/, /IH/, and /UH/ corresponding to the Arabic diacritical marks Fatha, Damma, and Kasra respectively. The /AA/ is the pharyngealized allophone of /AE/, which appears after an emphatic letter. Similarly, the /IX/ and /UX/ are the pharyngealized allophones of /IH/ and /UH/ respectively. When /AE/ appears before an emphatic letter, its allophone /AH/ is used instead. When a short vowel is located between two nasal letters in the same syllable it is likely to be nasalized. The allophones /AN/, /IN/, and /UN/ are the nasalized versions of /AE/, /IH/, and /UH/ respectively.

The regular Arabic long vowel allophones are /AE:/, /IY/ and /UW/ respectively. The length of a long vowel is normally equal to two short vowels. The allophones /AY/ and /AW/ are actually two vowel sounds in which the articulators move from one post to another. These vowels are called Diphthongs. The allophone /AY/ appears when a Fatha comes before an undiacritized Yeh. Similarly, /AW/ appears when a Fatha comes before an undiacritized Waw.

The Arabic voiced stops phonemes /B/ and /D/ are similar to their English counter parts. /DD/ corresponds to the sound of the Arabic Dhad letter.

The Arabic voiceless stops /T/ and /K/ are basically similar to their English counter parts.

The sound of the Arabic emphatic letter Qaf is represented by the phone /Q/. The Hamza plosive sound is represented by the phone /E/, and the sound of Jeem (in many dialects) is represented by /G/.

The voiceless fricatives are produced with no vibration of the voice cords. The sound is produced by the turbulence flow of air through a constriction. The Arabic voiceless fricatives /F/, /S/, /TH/, /SH/, and /H/ are basically similar to their English twins. In addition, the Arabic phones /SS/, /HH/, and /KH/ are the sounds of the Arabic letters Sad, Hah, and Khah respectively.

Voiced Fricatives are generated with simultaneous vibration of the vocal cords. The Arabic voiced fricative phones are /AI/, /GH/, /Z/, and /DH/ corresponding to the sound of the Arabic letters: Ain, Ghain, Zain, and Thal.

The Arabic affricative sound /JH/ is similar to the corresponding one in English, while /ZH/ is a concatenation of a voiced stop followed by a fricative sound.

The Arabic resonants are similar to the the English resonant phones. These are /Y/ for Yeh, /W/ for Waw, /L/ for Lam, and /R/ for Reh.

3. Arabic Phonetic Dictionary

Using the selected phoneme set, we developed a set of rules that are used to automatically generate the phonetic pronunciations for Arabic words. We also created a set of tools that process the given Arabic text and generate all possible pronunciations for every word in the text.

Rules are provided for each Arabic letter available in the Unicode listing (45 letters). Each rule tries to match certain conditions on the context of the letter and provide a replacement from the phoneme list. Replacements can be one or more phonemes. Some letters don't have an effect on pronunciation or, depending on context, they might not be pronounced; in this case, the replacement will be empty.

Each rule follows this format:

(pre_condition) . (post_condition) -> replacement

The left hand side of the rule is a PERL-like regular expression with the following definitions:

Each letter in the Arabic alphabet is referenced by its name as defined in the Unicode standard.

The dot (.) in the middle marks the current position (which is also the current letter) in the word.

Multiple classes are defined to simplify the rules syntax. Each class is referenced by its symbol (L, D, S, etc.) surrounded by angle brackets (< >). The classes are:

<L>: All Arabic consonants. <D>: Diacritic marks (Fathatan, Dammatan, Kasratan Fatha, Damma, Kasra, Shadda, and Sukun). <S>: Word Start. <T>: Word End. <SH>: Shamsi Letters (Teh, Theh, Dal, Thal, Reh, Zain, Seen Sheen, Sad, Dad, Tah, Zah, Lam, and Noon). <V>: Vowels (Fatha, Damma, Kasra, and Shadda). <VA>: Vowels without Shadda (Fatha, Damma, and Kasra). <P>: Prefix letters (Waw, Beh, Feh, Kaf, and Lam). <E>: Emphatic letters (Tah, Sad, Dad, and Zah). <PH>: Pharyngeal letters (Qaf, Ghain, Khah, and Reh).

The pre-condition has one of the following formats: (?<=pattern): context before the current position matches the pattern. (?<!pattern): context before the current position does not match the pattern.

In the same way, the post-condition has one of the following formats: (?=pattern): context after the current position matches the pattern. (?!pattern): context after the current position does not match the pattern.

Patterns use the following operators to define expressions:

Alternation: A vertical bar (|) is used to separate alternatives. **Grouping:** Parentheses () are used to define groups that determine scope and precedence of the operators and build complex expressions. **Optional matching:** A question mark (?) is used to mark parts of the expression that may or may not exist.

The right hand side of the rule defines the replacement, which can either be a phoneme or a sequence of phonemes from the phoneme list, or the letter might not have a matching phoneme and will be omitted from pronunciation. This case is marked with an asterisk (*) on the right hand side.

We define a rule set that covers all possible Arabic letters that are used in typing. Many of the rules are straight forward; they match the Arabic letters to their corresponding phonemes as explained in Table 1. Vowels require more elaborate rules to cover all possibilities. Special attention is required for nasalized consonants (Meem and Noon) and a few more exceptions that will be explained in the following sections.

3.1 Consonants

All Arabic consonants have a directly matching phoneme. However, some letters (Theh, Jeem, Thal, Zah, and Qaf) have multiple possible pronunciations that are due to dialectical differences.

A Noon followed by a Beh is usually converted to a Meem. This rule is optional and the speaker might not follow.

```
NOON:
.(?=BEH)-> M
.-> N
```

If the letter Dal is followed by a vowelled Teh then it is omitted in pronunciation. A similar case applies to the letter Dad.

```
DAL:
.(?=TEH<V>)-> *
.(?!TEH<V>)-> D
```

3.2 The letter Lam

The letter Lam has a set of complex rules when it comes in a combination known as Al-Alta'rif (ـل). If that combination is followed by a letter from the Shamsi group then the Lam is not pronounced. This rule, however, is not mandatory.

```
LAM:
(?<=<P><V>)?ALEF FATHA?.(?= <SH>)-> *
```

3.3 Semi-vowels

The letter Waw is sometimes treated as semi-consonants /W/ or /AW/ and other times it is treated as a long vowel, depending on its context. If the letter Waw is not vowelled and is preceded by a Damma then it is considered to be a long vowel.

The case of the semi-vowel /AW/ is similar to the long vowel, except it is then preceded by a Fatha. In this case the Waw is omitted. The insertion of the /AW/ phoneme is handled by the Fatha rules as it will follow shortly.

In the rest of the cases the Waw is converted to the semi-vowel /W/.

WAW:

```
(?<=<FATHA|DAMMA>).(?!<V>)-> *  
(?<=<FATHA|DAMMA>).(?!<V>)-> W  
(?!<FATHA|DAMMA>)-> W
```

The letter Yeh follows a similar pattern.

3.4 Tanween

The rules for this group differentiate between the emphatic and or pharyngeal versions of the vowels. Each rule appends an /N/ sound to the pronunciation.

```
FATHATAN:  
(?!<E>|<PH>)-> AE N  
(?<=<E>)-> AH N  
(?<=<PH>)-> AA N  
DAMMATAN:  
(?!<PH>)-> UH N  
(?<=<PH>)-> UX N  
KASRATAN:  
(?!<PH>)-> IH N  
(?<=<PH>)-> IX N
```

3.5 Vowels

Vowels have many versions. They are either short or long vowels. Both short and long vowels also can be normal ones or either emphatic or pharyngeal, depending on the surrounding letters. We developed rules that take care of all these situations. For instance, the rules for Fatha are:

```
FATHA:  
(?!<E>|<PH>).(?!ALEF((WAW|YEH)(<L>|<T>)))-> AE  
(?!<E>|<PH>).(?!ALEF)-> AE:  
(?<=<ALEF_WITH_MADDA_ABOVE>)-> AE:  
(?<=<E>).(?!ALEF((WAW|YEH)(<L>|<T>)))-> AH  
(?<=<E>).(?!ALEF)-> AH:  
(?<=<PH>).(?!ALEF((WAW|YEH)(<L>|<T>)))-> AA  
(?<=<PH>).(?!ALEF)-> AA:  
(?!WAW (<L>|<T>))-> AW  
(?!YEH (<L>|<T>))-> AY
```

The first two rules are responsible for the long and short versions of the normal vowels. The third rule is also for long vowels where the Fatha is followed by an Alef with Madda Above.

Rules 4-7 are for the emphatic and pharyngeal versions of the vowel.

Rules 8-9 takes care of the semi-vowels /AW/ and /AY/ as mentioned in the rules for the Waw and Yeh.

Rules for Damma and Kasra follow the same logic for the FATHA.

4. Arabic Broadcast News Corpus

Testing the rules for generation of the Arabic pronunciation dictionary is performed on an actual Arabic Automatic Speech Recognition (ASR) system. The ASR system was trained using an Arabic broadcast news corpus. The corpus consists of a total of 249 news stories, summing up to 5.4 hours of speech. The recordings were then split into 4572

utterance files with an average file length of 4.5 seconds. The vocabulary list contains 14,231 words.

5. Evaluation of the Arabic Pronunciation Dictionary

To test and validate the proposed phoneme set and the conversion rules we split the audio recordings into training and testing sets. The training set contained around 4.3 hours of audio while the testing set contained the remaining 1 hour. We used the CMU language toolkit to build a statistical language model from the transcription of the full 5.4 hours of audio.

Next, several test cases were built to validate our choice of the phoneme set. The main focus was on vowels since they impose most of the complexity and variety to the rules. For each of these cases the entire Arabic speech recognition engine is re-trained with the modified dictionary, and performance of the speech recognition is evaluated on the test set of the speech utterances (1144 voice files).

First, we tested the ASR system using the phone set and the rules outlines in the previous sections as the base system. We then developed four test cases.

The first test case studies the effect of removing the emphatic and pharyngeal vowels. The test case was built in which we removed these vowels from the phoneme set and omitted the rules for emphatic and pharyngeal vowels.

In the second test case we examined the effect of merging the long and short versions of vowels into one vowel (for example /AE/ and /AE:/) to test whether the tri-phone models in the speech recognition engine would be capable of handling these vowels without the need to introduce additional phonemes.

In the third case we examined alternative rules for handling the vowels preceding the definite article AL ALTA'REEF (ال). And for certain cases of the short vowel /AE/ and the long vowel /AE:/.

In the fourth case we examined alternative rules for dealing with gemination Shaddah, and for co-articulation effects of the emphatic consonants on the preceding vowels.

Results for these test cases are shown in Table 2.

Further analysis indicates that many of the word substitution errors are due to slight differences (deletion/substitution) of diacritical marks, especially the end cases. Since MSA text is written without diacritical marks, the error analysis was carried out once more after removing all the diacritical marks. The percentage of the correctly recognized words was 92.84%. The WER dropped to 9.0%.

The results of these tests lead us to conclude that it is necessary to include both emphatic and pharyngeal vowels and maintain separate phonemes for short and long vowels. The tests clearly validate the proposed phoneme set and the proposed rules for automatic generation of the Arabic pronunciation dictionaries for Arabic speech recognition applications. However, we

believe also that more research work is still needed to achieve better accuracy results.

Table 2. Summary of the performance of the AASR system for different phone/rules test cases.

	Test case	accuracy	I	D	S	WER
	Base system	90.1	168	82	838	11.71
1-	Emphatic and pharyngeal versions of vowels were removed	89.8	168	89	858	12
2-	Long versions of Vowels were removed	87.96	179	99	872	12.33
3-	Alternative rules for FATHA and for /AE/ and /AE:/ in the definite articles	88.47	214	80	991	13.84
4-	Alternative rules for co-articulation effects of the emphatic consonants	89.81	157	77	869	11.88

6. Conclusion

The paper provides a comprehensive set of rules for automatic generation of Arabic phonetic dictionary. This result was part of an on-going research towards achieving large vocabulary, speaker independent, natural Arabic automatic speech recognition system. The generated dictionary was based on about 14 K vocabulary in 5.4 hours of Arabic broadcast news corpus. The Dictionary was tested in a large vocabulary speaker independent Arabic speech recognition system. The speech recognition system achieves a comparable accuracy to the English ASR system for the same vocabulary size. Further enhancement will be carried out during the next phase of this research work, including extending the corpus to 40 hours, enhancing the rule based phonetic dictionary, and using a finer parameterization of the acoustic model.

7. Acknowledgments

This work was supported by a grant #AT-24-94 from King Abdulaziz City of Science and Technology. The authors would like also to thank King Fahd University of Petroleum and Minerals for its support in carrying out this project.

8. References

[1] X.Huang, A. Acero, and H. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.
 [2] Young, S. (1996), "A review of large-vocabulary continuous-speech recognition", *IEEE Signal Processing*

Magazine, pages 45-57, 1996.

[3] Carnegie Mellon University. CMU pronouncing dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

[4] M. Elshafei Ahmed, "Toward an Arabic Text-to-Speech System", *The Arabian Journal of Science and Engineering*, Vol. 16, No. 4B, pp.565-583, 1991.

[5] Mansour Alghamdi, Husni Almuhtasib, Mustafa Elshafei, "Arabic Phonological Rules", *King Saud University Journal: Computer Sciences and Information*. Vol. 16, pp. 1-25, 2004.

[6] K.F. Lee, "Large Vocabulary Speaker- Independent Continuous Speech Recognition: The SPHINX System," PhD Thesis, Carnegie Mellon University, 1988.

[7] HTK speech recognition tool kit. <http://htk.eng.cam.ac.uk/>

[8] CMU Sphinx Group. <http://www.speech.cs.cmu.edu/sphinx/>

[9] Algamdi, Mansour, "KACST Arabic Phonetics Database", *The Fifteenth International Congress of Phonetics Science*, Barcelona, 3109-3112, 2003.

[10] Moustafa Elshafei, Husni Almuhtasib and Mansour Alghamdi, "Techniques for High Quality Text-to-speech", *Information Science*, 140 (3-4) 255-267, 2002.

[11] Mohamed Ali, Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Al-Ghamdi, "Automatic Segmentation of Arabic Speech", Workshop on Information Technology and Islamic Sciences, Imam Mohammad Ben Saud University, Riyadh, March 2007.

[12] Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Alghamdi, "Machine Generation of Arabic Diacritical Marks", *Proceedings of the 2006 International Conference on Machine Learning; Models, Technologies, and Applications (MLMTA'06)*, June 2006, USA.

[13] H. Satori, M. Harti, N. Chenfour, "Introduction to Arabic Speech Recognition Using CMU Sphinx System", *Information and Communication Technologies International Symposium proceeding ICTIS07*, 2007.

[14] Hussein A.R. Hiyassat, Automatic Pronunciation Dictionary Toolkit for Arabic Speech Recognition Using SPHINX Engine, Ph.D., Arab Academy for Banking and Financial Sciences, Amman, Jordan, 2007.

[15] K. Kirchhofl, J.Bilmes, S. Das, N. Duta, M. Egan, G. Ji, F. He, J. Henderson, D. Liu, M. Noamany, P. Schoner, R. Schwartz, and D. Vergyri, "Novel Approaches to Arabic Speech Recognition: Report from the 2002 John-Hopkins Summer Workshop", *ICASSP 2003*, pp. 1-344-1347, 2003.