Study on pharyngeal and uvular consonants in foreign accented Arabic for ASR

Yousef Ajami Alotaibi, Ghulam Muhammad *

Department of Computer Engineering, College of Computer and Information Sciences, King Saud University,
P.O. Box 51178, Riyadh 11543, Saudi Arabia

Received 6 February 2008; received in revised form 15 April 2009; accepted 27 April 2009
Available online 15 May 2009

Abstract

This paper investigates the unique pharyngeal and uvular consonants of Arabic from the point of view of automatic speech recognition (ASR). Comparisons of the recognition error rates for these phonemes are analyzed in five experiments that involve different combinations of native and non-native Arabic speakers. The most three confusing consonants for every investigated consonant are discussed. All experiments use the Hidden Markov Model Toolkit (HTK) and the Language Data Consortium (LDC) WestPoint Modern Standard Arabic (MSA) database. Results confirm that these Arabic distinct consonants are a major source of difficulty for Arabic ASR. While the recognition rate for certain of these unique consonants such as /h/ can drop below 35% when uttered by non-native speakers, there is advantage to include non-native speakers in ASR. Besides, regional differences in pronunciation of MSA by native Arabic speakers require the attention of Arabic ASR research.

© 2009 Elsevier Ltd. All rights reserved.

Keywords: Arabic; Foreign accents; HMMs; Pharyngeal; Uvular; Speech recognition

1. Introduction and background

Arabic is a Semitic language which has many differences when compared with Indo-European languages such as English. Some of the differences include unique phonemes and phonetic features, and a complicated morphological word structure. It has been shown that major difficulties in automatic speech recognition (ASR) systems dedicated to Modern Standard Arabic (MSA) can be attributed to distinctive characteristics of the Arabic sound system, namely, geminate, emphatic, uvular, and pharyngeal consonants, and vowel duration (Selouania and Caelen, 1998; United Nations, 2003).

Compared to other languages, Arabic ASR has been the subject of a relatively small amount of research. Most efforts have concentrated on developing recognizers for MSA, which is the formal linguistic standard used throughout Arabic-speaking countries in the media, lectures, courtrooms (Kirchhoff et al., 2003). The
present paper concentrates on the analysis and investigation of five Arabic unique pharyngeal and uvular sounds from an ASR perspective. This investigation also focuses on the effect of foreign-accented pronunciation on accuracies in the ASR system.

This first section provides background for this research, and explains related topics that will give readers an overview of some of the difficulties that Arabic ASR faces, including those with unique Arabic consonants.

1.1. Arabic language

Arabic is a Semitic language, and it is one of the world’s oldest languages. Currently, it is the fifth most widely spoken language in the world. The estimated number of Arabic speakers is 250 million, of whom roughly 195 million are first-language speakers and 55 million are second-language speakers. Since it is also the language of religious instruction in Islam, many more speakers have at least a passive knowledge of the language. Arabic is an official language in more than 22 countries (Kirchhoff et al., 2002). It is the “first” language in countries such as Saudi Arabia, Jordan, Oman, Yemen, Egypt, Syria, and Lebanon (Alkhouli, 1990; Al-Zabibi, 1990).

Compared to MSA, Classical Arabic is an older, literary form of language, exemplified by the type of Arabic used in the holy Quran. Spoken Arabic is a collection of regional and national varieties that are derived from Classical Arabic. Arabic dialects are primarily oral languages; written material is almost invariably in MSA. As a result, there is a serious lack of language model (LM) training material for dialectal speech. MSA is a version of Classical Arabic with a modernized vocabulary (El-Imam, 1989), and it is a formal standard common to all Arabic-speaking countries. It is the language used in the media (television, radio, press, etc.), in official speeches, in universities and schools, and generally speaking, in any kind of formal communication situation (Kirchhoff et al., 2002).

The syllable types that are allowed in MSA Arabic are CV, CVC, and CVCC, where V indicates a (long or short) vowel and C indicates a consonant. Arabic utterances must start with a consonant (Alkhouli, 1990), and all Arabic syllables must contain at least one vowel. In addition, while Arabic vowels cannot occur in word-initial position, they can occur between two consonants or in word-final position. This is in contrast with other major languages, like English, Japanese, etc. In Japanese language, vowel can occur at any position of a word and most of the Japanese words end with vowel like pronunciation. Arabic syllables can be classified as short or long. The CV syllable type is a short syllable while all others are long. Syllables can also be classified as open or closed; an open syllable ends with a vowel, while a closed syllable ends with a consonant. For Arabic, a vowel always forms a syllable nucleus, and there are as many syllables in a word as there are vowels in it (El-Imam, 1989).

The Arabic alphabet consists of 29 letters, 26 of which represent consonants. The remaining three letters represents the long vowels of Arabic (the phonemes /iː/, /aː/, /uː/) and, where applicable, the corresponding semivowels (the phonemes /j/ and /w/). Table 1 shows all Arabic alphabet letters and their correspondences to consonant and semivowel phonemes. This table also shows the phonetic description of each phoneme including the place of articulation. Throughout this paper we use the symbols of International Phonetic Alphabet (IPA) and symbols of Language Data Consortium (LDC) WestPoint Modern MSA database phoneme. This is decided because we are using WestPoint Corpus in our research and they are using their custom Arabic phoneme list.

1.2. Phonology and morphology

A phoneme is the smallest unit of sound that corresponds to an element of human speech that can indicate differences in meaning between words or sentences. Phonemes are often classified into two major groups: vowels and consonants. In terms of their phonetic realization, vowels contain no major airflow restriction in the vocal tract; consonants involve a significant restriction of airflow and are therefore weaker in amplitude and often noisier than vowels (Rabiner and Juang, 1993; Deller et al., 1993). Arabic has 36 phonemes consisting of three short vowels (/i/, /a/, /u/), three long vowels (/iː/, /aː/, /uː/ which are the counterparts of the short vowels), and 28 consonants (Alghamdi, 2004).

Arabic has noticeably fewer vowels than other major languages. For example, some varieties of American English have at least 12 vowels and Japanese has not less than five vowels. In Bengali, which is the seventh
most spoken language in the world, there are 11 vowels. On the other hand, Arabic has three long and three short vowels (Deller et al., 1993). In addition, vowel lengthening in Arabic is phonemic. Also, in Arabic, there are diacritics which play the role of vowels. Some Arabic dialects may have additional or fewer consonant phonemes. For example, Egyptian Arabic lacks the phonemes /b/ and /d/, and it replaces /z/ with /g/ (Kirschhoff et al., 2002). On the other hand, these dialects may have different vowels. For instance, Levantine dialect has at least two extra type of diphthongs /aj/ and /aw/. Also Egyptian dialect has other extra vowels (Elgendy, 2001). Arabic phonemes contain two distinctive classes that are named pharyngeal and emphatic phonemes. These two classes are primarily found in Semitic languages like Hebrew and Arabic (Alkhouli, 1990; Elshafei, 1991). Delattre (1971) found pharyngeal features in other languages like German, Spanish as well.

### 1.3. Unique consonants in Arabic

There are some unique consonants in many languages. For example, in Japanese, there are /ʃ/ and /n/, which are very short in pronunciation and difficult to recognize (Ghulam et al., 2005). In Bengali, there exist /f/, /g/, and /d/, which are unique in that language. All Slavic languages have palatalized consonants like /v/, /p/, /f/, /m/ and /b/. Armenian language has soft voiceless consonant /tʃ/ which can be sound as close to /d/.

Similarly, in Arabic, some unique consonants exist that cannot be found in other languages. For example, there are four emphatic consonants in Arabic: two plosives, /d/ and /t/, and two fricatives, /s/ and /ʃ/ (Al-Muhtaseb et al., 2000; Ouni et al., 2005; Selouania and Caelen, 1998). /d/ is a voiced emphatic stop with an alveo-dental point of articulation. As this phone is rare in human languages, Arabic is commonly called “The Dhaad language”, where Dhaad is the name of the spoken Arabic letter that carries the /d/ phoneme. /t/ is an unvoiced emphatic stop with an alveo-dental point of articulation. /s/ is an unvoiced emphatic fricative with an alveo-dental point of articulation. Finally, /ʃ/ is a voiced emphatic fricative with an inter-dental point of articulation (Alkhouli, 1990).
Also there are two pharyngeal phonemes in Arabic: fricatives, /\?/ and /h/ (Alghamdi, 2004; Alkhouli, 1990), and three uvular phonemes: one stop /q/, two fricatives /\?/ and /x/ – that are part of interested phonemes in our research. There is some controversy among researchers about the place of articulation of these unique phonemes in Arabic. Table 2 shows the two pharyngeal and three Arabic uvular sounds. This table displays the disagreement among selected group of Arabic phoneticians about the place of articulation of these five Arabic phonemes. More description about pharyngeal and uvular phonemes will be represented in the results section.

Newman (2002) investigated the phonetic status of Arabic within the world’s languages, with the concentration on the uniqueness of special Arabic phonemes. In this study he considered the framework of IPA as used with the UCLA Phonological Segment Inventory Database – commonly known as UPSID. In that database, there are 317 languages with total of 58 phonetic features. By concentrating on only pharyngeal and uvular Arabic phonemes that are of interest in our paper, he concluded that the voiced pharyngeal fricative /\?/ is limited only to eight languages (2.5%) – five of them are Afro-Asiatic, where the lengthened version is unique to Arabic. He reported that this phoneme was seen as a fricative in the high classical style of Quraan recitation, whereas in formal MSA it tends to be a stop, depending on the linguistic background of the speaker. The voiced uvular fricative /\?/ is reported in only 14 languages (4.38%) while the long version of this phoneme is unique for Arabic. In addition to this, Newman found that the pharyngeal unvoiced phoneme, /\?/, is accruing in only 13 languages (4.1%), and its longer variant was found in Arabic and Shilha. On the other hand, he attributed the uvular plosive /q/ as the least stable sound in as much as in many local varieties of Arabic. To make it clear, /q/ phoneme is realized either as a voiced velar plosive [g] (e.g., in some regions of Egypt and few regions of Tunisia), or as glottal stop as the case in the Syro-Lebanese and Cairene. This phoneme was reported in only 38 languages (11.9%) in UPSID. Finally in that study, the unvoiced uvular fricative /x/ occurs in only 27 languages (8.5%) within the database (interestingly enough from nearly every continent), whereas the lengthened variety is limited to only Arabic and two more other languages.

1.4. Arabic speech processing

Although digital Arabic speech processing is still in its infancy compared to languages such as English or Japanese, there have been several advances in this area of research. Kirchhoff et al. (2003) worked on a novel approach to Arabic ASR by concentrating on problems such as the absence of short vowels and other pronunciation information in Arabic text, the morphological complexity of Arabic, and discrepancies between diacritical and formal Arabic. They used LDC’s CallHome Arabic Speech Corpus. Their research produced three main outcomes. First, they showed that using phonetic information available in the form of romanized as opposed to vowelless transcriptions significantly improves word error rate; indeed, it is possible to obtain improvements by using automatically romanized data. Second, they observed an improvement by using morphologically based LMs. Finally, they found that various methods of using MSA text data to improve the CallHome LM did not yield any improvement.

Selouania and Caelen (1998) designed a mixture of artificial neural network experts for automatically recognizing Arabic consonants, including the four emphatic consonants of Arabic. Their system used time delay
neural networks and an autoregressive backpropagation algorithm (AR-TDNN). They used perceptual linear predictive coefficients, energy zero crossing rate and their derivatives as the features extracted from their front-end processor. They observed an error rate of 14.7% for the emphatic consonants. In the case of the best of the three systems, the one based on a parallel structure of neural network experts, they noted a failure in identifying the emphatic /dʒ/ consonant. Their explanation is that the problem does not reside in difficulties inherent to the consonant’s acoustic properties but rather in the poor ability of speakers, including native speakers, to pronounce it correctly. Their overall results showed that all designed systems had relatively high error rates for emphatic consonants when compared to fricatives, plosives, nasals, and liquid consonants.

Nureddeen (2008) investigated speech act performance in Sudanese Arabic. Her study was to outline the pragmatic rules that govern the use of language in different cultures and to show how findings can be used to facilitate communication between people from different socio-cultural backgrounds.

1.5. Non-native language ASR

There have been a significant amount of works done on ASR with non-native speakers in some of the major languages like English, German, Japanese, Chinese, etc. Tomokiyo (2000) reported linguistic properties of non-native speech, where the Japanese and the Chinese speakers were to speak English. She found that (i) the overall speaking rate for non-native speakers was 2/3 that of native speakers, (ii) the silence rate was double that of native speakers, and (iii) average phoneme length for non-native speakers was 1.5 times that of native speakers. These differences of non-native speech from native speech affected the performance of ASR. Uebler (2001) presented approaches to speech recognition in seven languages. The languages were German (two instances: one with native and non-native speakers, and the other with native speakers only), Italian, Slovak, Slovenian, Czech, Japanese and English. He concluded that for cross-lingual recognition, the choice of language for training the recognizer was important for the performance. He also found that performance increased if training data of more languages were involved and thus both acoustic units were modeled with more variety and more training material as well as more different acoustic units were modeled overall.

Aljasser (2008) revealed that phonetically training should form part of second language ear training and pronunciation programs in one of his experiments. In that experiment native English speakers were asked to recognize English from Arabic speakers. Awais et al. (2007) developed concurrent neural networks to identify Arabic phoneme spoken by nonnative Arabic speakers. Alotaibi et al. (2008) work concerned with the problem of non-native speech in a speaker independent, large-vocabulary speech recognition system for MSA Arabic. They analyzed some major differences at the phonetic level in order to determine which phonemes have a significant part in the recognition performance for both native and non-native speakers. The performance of an HMM-based Arabic speech recognition system was analyzed with respect to speaker gender and its native origin. The WestPoint Modern Standard Arabic database from the Language Data Consortium (LDC) and the Hidden Markov Model Toolkit (HTK) are used throughout all experiments. Their study showed that the best performance in the overall phoneme recognition is obtained when non-native speakers are involved in both training and testing phases.

2. Experimental framework

The system presented in this paper is designed to recognize Arabic phonemes. In this investigation we analyze the performance of the system with respect to the pharyngeal consonants – /s/ and /h/ – and the uvular consonants /y/, /q/, and /x/. The study focuses on the effect of native and non-native speakers in both training and testing data. The accuracies with respect to all five consonants and in all conducted experiments are reported and investigated in detail. The effect of the mother tongue on pronouncing these pharyngeal and uvular consonants was investigated in this research paper. It is well-known that these sounds are very difficult to pronounce by non-native Arabic speakers. In addition to this, the effect was analyzed in the performance of ASR systems when dealing with these Arabic sounds. This means that we are considering the set of Arabic consonants, the ASR system, and the Arabic mother tongue as the main components of this research.
2.1. ASR method and platform

Hidden Markov Models (HMMs) are a well-known and widely-used statistical method for characterizing the spectral features of speech frame. The assumption underlying HMMs is that the speech signal can be well characterized as a parametric random access, and the parameters of the stochastic process can be predicted in a precise, well-defined manner. HMMs provide a natural and highly reliable way of recognizing speech for a wide range of applications (Juang and Rabiner, 1991; Rabiner, 1989). The Hidden Markov Model Toolkit (HTK) (Young et al., 2005) is a portable toolkit for building and manipulating HMMs; it is widely used for designing, testing, and implementing ASR systems and related research tasks. HTK was used in all experiments reported here.

2.2. Database

In our experiments, we used the WestPoint Arabic Speech Corpus provided by LDC (LDC, 2002). The corpus was timeless labeled automatically forced aligned. We used this corpus as given by LDC and we did not analyze the errors of incorrect labeling or the existence of other kinds of errors. In addition to this, no system errors verification against the corpus in order to search for possible database mistakes. This corpus consists of collections of four main Arabic scripts. Collection Script 1 contains 155 sentences, uttered by all 74 native speakers of Arabic. Script 1 has a total of 1152 tokens and 724 types. Collection Script 2 contains 40 sentences used by 23 of the non-native speakers. Script 2 has a total of 150 tokens and 124 types. Collection Script 3 is comprised of 41 sentences used by four of the non-native speakers. It has a total of 138 tokens and 84 types. Finally, there is Collection Script 4, which contains 22 sentences used by nine of the non-native speakers, all of them third-year Arabic speakers. It has a total of 72 tokens and 59 types; the total number of distinct words is 1131 Arabic words. All scripts were written with MSA as the target language and are diacritized.

A descriptive summary of this database is given in Table 3. As shown in this table, the amount of data provided by the native speakers of Arabic is significantly greater than that provided by the non-native speakers. From the documentation provided by LDC, it appears that all members of the non-native Arabic speaker group are native speakers of English. The corpus includes both male and female speakers. We used this corpus because this is the only available public corpus that includes native and non-native MSA Arabic speakers.

<table>
<thead>
<tr>
<th>LDC WestPoint Corpus summary.</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of speakers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>41</td>
<td>34</td>
<td>75</td>
</tr>
<tr>
<td>Non-native</td>
<td>25</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>Totals</td>
<td>66</td>
<td>44</td>
<td>110</td>
</tr>
<tr>
<td><strong>Hours of data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>6</td>
<td>4.4</td>
<td>10.4</td>
</tr>
<tr>
<td>Non-native</td>
<td>0.74</td>
<td>0.28</td>
<td>1.02</td>
</tr>
<tr>
<td>Totals</td>
<td>6.74</td>
<td>4.68</td>
<td>11.42</td>
</tr>
<tr>
<td><strong>Megabyte of data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>913</td>
<td>663</td>
<td>1576</td>
</tr>
<tr>
<td>Non-native</td>
<td>111</td>
<td>42.4</td>
<td>153.4</td>
</tr>
<tr>
<td>Totals</td>
<td>1024</td>
<td>705.4</td>
<td>1729.4</td>
</tr>
<tr>
<td><strong>Number of speech files</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>4107</td>
<td>3163</td>
<td>7270</td>
</tr>
<tr>
<td>Non-native</td>
<td>883</td>
<td>363</td>
<td>1246</td>
</tr>
<tr>
<td>Totals</td>
<td>4990</td>
<td>3526</td>
<td>8516</td>
</tr>
</tbody>
</table>
2.3. System description and parameters

A complete ASR system based on HMMs was developed to carry out the goals of this research. This system was partitioned into three modules according to their functionality. First is the training module, whose function is to create the knowledge about the speech and language to be used in the system. Second is the HMM bank, whose function is to store and organize the system knowledge gained by the first module. Finally, there is a recognition module whose function is to try to figure out the meaning of the input speech given in the testing phase. This was done with the aid of the HMM models mentioned above.

As given in Table 4, the parameters of the system were 22 kHz sampling rate with 16 bit sample resolution, 25 ms Hamming window duration with a step size of 10 ms, MFCC coefficients with 22 as the length of cepstral liftering and 26 filter bank channels, 12 as the number of MFCC coefficients, and 0.95 as the pre-emphasis coefficients.

Phoneme-based models are good at capturing phonetic details. Context-dependent phoneme models are widely used to characterize formant transition information, which is very important for the discrimination of confusable phonemes. Our baseline system is designed as a phoneme-level recognizer with 3-state, continuous, left-to-right, no skip HMM models.

The baseline system considers all 37 MSA monophones as given in the LDC catalog (2002). We note that the WestPoint Corpus contains more monophones than the number of MSA phonemes mentioned in the linguistic literature (Alkhouli, 1990; Elshafei, 1991; Omar, 1991). Specifically, WestPoint has added three more phonemes: /g/ “voiced velar stop”, /aw/ “back upgliding diphthong”, and /ey/ “upper mid front diphthong”. In fact, the phoneme /g/ does not exist in MSA. We believe that the LDC used it because some native and non-native speakers produced it in certain MSA words. On the other hand, we believe that the two extra diphthongs were added because of variations in the pronunciations of non-native speakers, who speak English and possibly other languages. These phonemes exist in English but not in MSA. In any case, we decided to retain the WestPoint Corpus phonemes, transcriptions, and other settings without any modification. We believe that our decision will help the standardization with other research efforts that are using the same corpus and that have goals similar to ours. This will ensure meaningful comparisons among different researchers’ results.

Since most of the words consisted of more than two phonemes, context-dependent triphone models were created from the monophone models. Before this, the monophones models were initialized and trained by the training data. This was done with more than one iteration and was repeated for triphones models. Within the training phase, the model was aligned and tied by using the decision tree method. The last step in the training phase was to re-estimate the HMM parameters using the Baum–Welch algorithm (Rabiner, 1989) three times.

3. Results

The results reported here are based on the outcomes of the Arabic ASR system described above. This system computed the accuracies of all Arabic phonemes without using any LM. Five experiments were carried out in this investigation. These experiments differ only in the type of the training and testing data sets. These experiments are labeled as N/N, N/NN, NN/N, NN/NN, and M/M. In the experiments, N/N indicates that
native Arabic speakers are used in both training and testing phases. Native Arabic speakers are used in training data while non-native Arabic speakers are used in testing data of the N/NN experiment. Regarding the NN/N experiment, non-native Arabic speakers are used in training data, while native Arabic speakers are used testing data. Without using any native Arabic speakers, non-native Arabic speakers are used in both training and testing data of the NN/NN experiment. Finally, in the M/M experiment, a mixture of native and non-native Arabic speakers was used in both training and testing data.

The results are presented in three subsections. The first subsection reports the accuracies for the pharyngeal consonants and draws some preliminary conclusions. Similarly, the second subsection presents the same for the uvular consonants. The last subsection is a general discussion based on the observations.

3.1. Pharyngeal consonants

Fig. 1 plots the accuracies of the two Arabic pharyngeal consonants, /\?/ and /\h/, for all five experiments. In addition to this, Table 5 depicts the three most confusing Arabic consonants for every pharyngeal and uvular phoneme and for every experiment. These two Arabic consonants are subset of the consonants that feature Arabic languages and they cannot be found in Latin languages such as English. This subsection gives a brief description about each consonant followed by discussion of the experiments outcomes.

![Pharyngeal Sounds](image)

**Fig. 1.** Accuracy of pharyngeal sounds for different experiments (legend: C stands for \? and H stands for \h).

<table>
<thead>
<tr>
<th>Arabic alphabet carrier</th>
<th>IPA</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N/N N/NN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NN/N NN/NN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M/M</td>
</tr>
<tr>
<td>Ain</td>
<td>/?</td>
<td>/?, r, l/</td>
</tr>
<tr>
<td>Ghain</td>
<td>/y/</td>
<td>/?, r, \d\j/</td>
</tr>
<tr>
<td>H/\aa</td>
<td>/h/</td>
<td>/?, r, l/</td>
</tr>
<tr>
<td>Qaaf</td>
<td>/q/</td>
<td>/?, r, l/</td>
</tr>
<tr>
<td>Khaa</td>
<td>/x/</td>
<td>/?, r, l/</td>
</tr>
</tbody>
</table>

Table 5
Three most confusing consonants with all studied consonants and for experiments.
3.1.1. Ain /ʔ/

This phoneme is one of the unique Arabic consonants that cannot be found in Latin languages such as English. All Arabic phoneticians agree about the place of articulation regarding this phoneme and they determine it as pharynx (Alghamdi, 2001; Alkhouli, 1990; Bisher, 1990; Borden and Harris, 1990; Harakat, 1998; Mahamoud, 2003; Nour-Aldeen, 1992; Omar, 1991). This phoneme can be described as fricative, voiced, non-emphatic, and pharyngeal sound. It was investigated in all the five experiments and the accuracies of the system for this specific phoneme are depicted in Fig. 1. As we can see from that figure, the second best accuracy for this phoneme was achieved when the system was trained and tested by only native Arabic speakers (i.e., in the experiment N/N). In this case the accuracy was 71.4%. On the other hand, the experiment NN/NN encountered the lowest system accuracy regarding this phoneme when non-native Arabic speakers’ data was used in both training and testing phases. This low accuracy was 50.6%. This suggests that there is inconsistency of pronouncing /ʔ/ between non-native Arabic speakers.

In the experiment M/M, where mixed native and non-native Arabic speakers were used in both training and testing phases, the system achieved the highest accuracy for this phoneme. This accuracy was 72.0%. In case of N/NN and NN/N, the system accuracy for /ʔ/ phoneme is medium compared to the best and worst cases mentioned above. Depending on Fig. 1, we can state a substantial conclusion about the robustness of using M/M over N/NN for both pharyngeal phoneme and the last phoneme in the uvular set as given in Fig. 2. This implied that the system is getting better immunity against missing these Arabic unique phonemes by mixing native and non-native speakers due to losing the actual acoustic features of these phonemes by avoiding non-native speakers alone. This result confirms the findings of Uebler (2001) and Alotaibi et al. (2008) described in Section 1.5.

By consulting Table 5, we can see /ʔ/ is mostly confused with /r/ phoneme in all experiments. This phoneme /r/ in Arabic language is trill, voiced, non-emphatic, and alveolar sound. /r/ phoneme also exists in English language but it is liquid sound instead of trill. This phoneme is vocalized as the Arabic version in WestPoint. In addition to that /ʔ/ phoneme was confused by /ɣ/, /ð/, and /l/ phonemes.

3.1.2. Hن/aa /h/.

The phoneme Hن/aa is an Arabic consonant that cannot be found in English phoneme list. All Arabic phoneticians agree about the place of articulation of this sound to be pharynx (Alghamdi, 2001; Alkhouli, 1990; Bisher, 1990; Borden and Harris, 1990; Harakat, 1998; Mahamoud, 2003; Nour-Aldeen, 1992; Omar, 1991). This phoneme can be described as fricative, unvoiced, non-emphatic, pharyngeal sound. This phoneme was
investigated in all the five mentioned experiments and its accuracy reached its highest value with the M/M experiment. The accuracy was 76.9%. It is a remarkable high accuracy for this group of Arabic phonemes for all experiments. This result implies that including native speakers in the training can significantly improve the robustness of the system. The worst accuracy was encountered in case of the experiment N/NN. The accuracy in this case was only 26% with 50% less than its above-mentioned maximum (i.e., in case of the M/M experiment). Indeed it is the worst of the worst accuracy in experiment for all Arabic pharyngeal and uvular sounds. This piece of information tells us that the non-native Arabic speakers pronounce this sound in a completely wrong manner.

Depending on our knowledge about the non-native Arabic speakers’ speech, non-Arab speakers have difficulty in perception and pronunciation of this sound and they very often convert it to /h/ phoneme. /h/ phoneme exists in most languages including Arabic and English and it is the glottal counterpart of the /h/ sound. Our experiments confirm this information because by referring to Table 5 we can see that /h/ phone was confused mostly by /h/ phoneme. This can be seen clearly in /h/ row of that table. This happened for all experiments but not N/NN experiment. This means that non-native Arabic speakers turn /h/ phoneme to its glottal counterpart, /h/. Other phonemes that caused confusions for /h/, depending on Table 5, are /r/, /ʔ/, and /ɣ/.

We can also observe that the accuracy of the /h/ phoneme even in the N/N experiment is below 60%. This is because /h/ is an unvoiced phoneme (see Table 1) that can heavily be affected by other neighboring phonemes.

3.2. Uvular consonants

Arabic language, in both the classical and MSA versions, has three uvular consonants, namely Ghain /ɣ/, Qaaf /q/, and Khaa /x/. Fig. 2 plots the accuracies of these three Arabic uvular consonants for all five experiments. Again, these three Arabic consonants are subset of the consonants that feature Arabic languages and they cannot be found in Latin languages such as English. This subsection gives a brief description about each consonant followed by discussion of the experiments outcomes.

3.2.1. Ghain /ɣ/

Some Arabic phoneticians described this Arabic phoneme as velar phoneme (Alkhouli, 1990; Omar, 1991); but others describe it as palatal one (Bisher, 1990; Borden and Harris, 1990). On the other hand, a third group of Arabic phoneticians describe it as uvular one (Alghamdi, 2001; Nour-Aldeen, 1992). The third group of authors is concentrating on Arabic MSA and avoiding regional dialects. In addition to this, quick comparative pronunciation to this sound will verify the correctness of this way of description. Also other groups who described the /ɣ/ as a non-uvular sound may be effected by regional dialects. In this paper, we are talking about MSA Arabic only without the affect of regional dialects. Also this confusion occurrence is influenced by the closeness in locations of the velar, palatal, and uvular in the human being vocal tract area. Some of the phoneticians mix these parts of the vocal tract in one part. Thus laying on the opinion of the later group of phoneticians we consider /ɣ/ phoneme a uvular phoneme in this paper. We say this because all /q/, /x/, and /ɣ/ phonemes have the same place of articulation when we experience vocalization of them in same setting of words. As a full description of this phoneme we can define it as a fricative, voiced, non-emphatic, and uvular Arabic phoneme.

/ɣ/ sound was investigated in all experiments conducted in our research where the best accuracies of the system regarding /ɣ/ phoneme were occurred in case of N/N and N/NN where the accuracies were 63.3% and 64%, respectively. The worst accuracy was encountered with experiment NN/N. /ɣ/ phoneme accuracy in this case was 47.5%. For all experiments, /r/ phoneme is one of the most confusing phonemes in recognizing /ɣ/ phoneme. This is a similar case in /s/ phoneme discussed above. /s/ and /ʔ/ phonemes are among those phonemes that disturb /ɣ/ phoneme accuracy in most of experiments. The information in /s/ and /ɣ/ rows of Table 5 suggests that there are big similarities in pronunciation of /s/, /ɣ/ and /r/.

3.2.2. Qaa /q/

One of the Arabic phoneticians described /q/ phoneme as pharyngeal phoneme (Alkhouli, 1990), but all others describe it as a uvular one (Alghamdi, 2001; Bisher, 1990; Borden and Harris, 1990; Harakat, 1998; Mahamoud, 2003; Nour-Aldeen, 1992; Omar, 1991). This phoneme can be described as stop, unvoiced,
non-emphatic, and uvular Arabic phoneme. In our experiment, we investigated /q/ phoneme and we found the best accuracy for /q/ in the experiment N/N. The performance of the system in recognizing /q/ phoneme was 55%. This performance rate is lower compared to the performances of the N/N experiment on other investigated phonemes. The result suggests that /q/ phoneme is a highly confusing phoneme even for native Arabic speech. Also, this is an unvoiced phoneme (Table 1) and it can easily be buried with other neighboring sound and noise. This phenomenon may affect the accuracy.

The worst performance for this phoneme, on the other hand, was encountered in the experiment NN/N. The accuracy of this phoneme in the mentioned experiment was 30.4%. Also the accuracy of this phoneme in experiment NN/NN was 31.9%. We can conclude that the non-native speakers cannot pronounce this phoneme correctly as can be shown from the accuracies of experiments NN/N and NN/NN. Depending on Table 5, this phoneme was confused with /ʔ/ phoneme in all experiment, which means that there is big similarity between /q/ and /ʔ/ phonemes. In fact /ʔ/ phoneme is the glottal counterpart of /q/ phoneme (i.e., /ʔ/ phoneme differ from /q/ phoneme only in the place of articulation where the former one is glottal and the latter one is uvular). /r/ and /ʕ/ are among the confusing phonemes of /q/ phoneme.

3.2.3. Khaa /x/

The /x/ phoneme is similar to /ɣ/ phoneme with regard to the phoneticians’ disagreement about its place of articulation. Some Arabic phoneticians described /x/ phoneme as velar phoneme (Alkhouli, 1990; Omar, 1991), but others describe it as palatal one (Bisher, 1990; Borden and Harris, 1990). On the other hand, a third group of Arabic phoneticians describe it as uvular one (Alghamdi, 2001; Nour-Aldeen, 1992). Again and similar to the case in /ɣ/ phoneme, we think the later description is more suitable to our work, where we consider MSA only without the influence of local dialects. Also other groups who described the /x/ as a non-uvular sound may be affected by regional dialects. Our judgment regarding the consideration of this description of the last group of authors is similar to that we did for Ghain sound above. Therefore, we consider /x/ phoneme a uvular phoneme in this paper. We say this because all /q/, /x/, and /ɣ/ phonemes have the same place of articulation when we experience vocalization of them in same kind of words.

Hence /x/ phoneme can be described as fricative, unvoiced, non-emphatic, and uvular sound. In our conducted experiments, this phoneme was investigated and it has been found that the best accuracy was encountered with experiment M/M. The accuracy for that case was 86.7%. This accuracy is excellent compared to all other experiments and phoneme accuracies in this group of sounds. This is an affirmation to the fact discovered by Uebler (2001) and Alotaibi et al. (2008) that non-native speakers in the training phase can increase the robustness of the system. The worst accuracy was encountered with experiment NN/N. The accuracy in that case was 34.5%. Similar low accuracy was encountered with experiment NN/NN. In case of experiment N/N the accuracy was 58%. Depending on Table 5, we can see that this phoneme was confused mostly by /h/, /ʔ/, and /r/ phonemes.

3.3. Discussion

Despite the fact that LDC WestPoint Arabic Speech Corpus, which is the only available corpus that includes native and non-native MSA Arabic speakers, has many disadvantages including selecting only non-native Arabic speakers who have English as their first language, not balanced speakers with respect to number of native/non-native, and problems in labeling and alignments, we found some interesting observations regarding native and non-native issue in Arabic speech. The accuracies in experiments N/N and M/M are, in general, better than other experiments namely N/NN, NN/N, and NN/NN. Depending on the outcomes of the experiments as shown in Figs. 1 and 2 we can notice that our system experiences difficulty in understanding the real acoustic features of Ain/H\ʕaa/Khaa/Ghain and also it faces difficulty in recognizing these sounds. This is obvious in the big improvement in case of M/M experiments in both sounds compared to using non-native (NN) subset alone in training and/or testing subsets. Mixing non-native speakers with native speakers in the training phase increases the robustness of a generalized system, where the system recognizes phones irrespective of native/non-native speakers. The upper ends of bars of Fig. 2 look like “U” shape for most studied phonemes. This means that the accuracy of the system for uvular phoneme in the experiments N/N and M/M are better than other experiments. In these two experiments, native Arabic
speakers were used in training and testing phases of the recognition systems. In the other three experiments that are located in the middle of the figure, non-native Arabic speakers were used in training phase and/or testing phase of the recognition system; hence the accuracies were relatively low. For pharyngeal phonemes, accuracies were relatively higher in the experiment NN/N comparing to the other two experiments mentioned just above. Some of Arabic native speakers who learn English for the first time face some difficulty in distinguishing both by ear and tongue between the phonemes /p/ and /b/ due to the absence of first one from the Arabic phoneme list. This is similar case as the one discovered by our system here. The variation in pronunciations of these phonemes (i.e., in this work) was considered in training phase of the NN/N experiment.

We have found that accuracies of some of the investigated phonemes, /h/, /q/ and /x/, even in the experiment N/N were below 60%. These phonemes are unvoiced and have relatively shorter time duration and less amplitude in time domain. Therefore, these unvoiced phonemes are highly confused by coarticulation effect giving low recognition accuracy.

We have another interesting observation from Table 5. We noticed that in N/N column of the table most confusing phonemes are /ʔ/ and /ɣ/ for all investigated five phonemes. In N/NN column of the table, most confusing phonemes are /ɾ/ and /ʔ/ for all investigated five phonemes. In NN/N column of the table, most confusing phonemes are /h/ for all investigated five phonemes. Finally, in M/M column of the table, most confusing phonemes are /tʃ/ and /ɾ/ for all investigated five phonemes. This implies that most disturbing phonemes for pharyngeal and uvular phonemes are dependent on the mother tongue of the speakers used in training phase and/or testing phase.

4. Conclusion

An Arabic phoneme recognition system was designed and used to investigate Arabic pharyngeal and uvular phonemes in Modern Standard Arabic (MSA). The investigation depended mainly on speech recognition outcomes. This speech recognition recognizes speech signal by using phoneme level without using any language model. The most three confusing phonemes that degraded accuracy of every phoneme in our set were presented and discussed. The most important outcome from all manipulation presented in this research can be summarized in that the non-native Arabic speakers induced a big portion of miss-recognizing pharyngeal and uvular sounds. This problem has emerged by wrong pronunciation to the five Arabic exclusive sounds. There are two reasons behind this problem: first, difficulty to pronounce correctly these sounds; second is the existence of other phonemes in Arabic phoneme inventory that have big similarity between them and the pharyngeal and uvular sounds under investigation. One example is the phoneme /h/ which is the glottal counterpart of /ɾ/ phoneme. Our research proved that /h/ phoneme badly confused with the /ɾ/ phoneme; thus increasing its error rate in the system output. Close aural inspection of the pronunciation by native Arabic speakers in the WestPoint Corpus found considerable regional variation, in other words a foreign-accented MSA. This suggests the need for research on Arabic ASR to control for regional and other social correlates of phonetic variation. Attention to processes such as the spread of the pharyngeal feature of the emphatic consonants to neighboring segments should also inform future work in this area. Also, the paper noted several significant directions for future research. Training the system with a mixture of native and non-native speakers can significantly improve the recognition accuracy of ASR system. Repeating this work by using a more comprehensive and larger in size corpus is one of these research topics. By considering a corpus that cover more non-native Arabic speakers who have their first languages other than English, more Arabic dialects from all overall the Arabic countries, phonetically rich scripts, conversational speech, and balanced speakers regarding gender, age, regions, and, culture. In addition to this, adapting the speech of non-native Arabic pronunciation of these specific phonemes is one of the untouched topics which need some digging. By using adaptation point of view we can consider adaptation of non-native Arabic speakers and incorrectly pronounced special Arabic phonemes.

References


Ouni, S., Cohen, M., Massaro, W., 2005. Training Baldi to be multilingual: a case study for an Arabic Badr. Speech Communication 45, 115–137.


