Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2008, Article ID 679831, 9 pages
doi:10.1155/2008/679831

Research Article

Experiments on Automatic Recognition of Nonnative Arabic Speech

Yousef Ajami Alotaibi,[1] Sid-Ahmed Selouani,[2] and Douglas O'Shaughnessy[3]

[1] *Computer Engineering Department, King Saud University, Riyadh 11451, Saudi Arabia*
[2] *Laboratoire de Recherche en Interactivité Homme Système LARIHS, Université de Moncton, Campus de Shippagan, New Brunswick, Canada E8S 1P6*
[3] *INRS-Energie-Matériaux-Télécommunications, Université du Québec, 800 de la Gauchetière Ouest, place Bonaventure, Montréal, Canada H5A 1K6*

Correspondence should be addressed to Sid-Ahmed Selouani, selouani@umcs.ca

Received 11 May 2007; Revised 5 October 2007; Accepted 13 January 2008

Recommended by Li Deng

The automatic recognition of foreign-accented Arabic speech is a challenging task since it involves a large number of nonnative accents. As well, the nonnative speech data available for training are generally insufficient. Moreover, as compared to other languages, the Arabic language has sparked a relatively small number of research efforts. In this paper, we are concerned with the problem of nonnative speech in a speaker independent, large-vocabulary speech recognition system for modern standard Arabic (MSA). We analyze 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 nonnative speakers. Special attention is given to specific Arabic phonemes. The performance of an HMM-based Arabic speech recognition system is 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. Our study shows that the best performance in the overall phoneme recognition is obtained when nonnative speakers are involved in both training and testing phases. This is not the case when a language model and phonetic lattice networks are incorporated in the system. At the phonetic level, the results show that female nonnative speakers perform better than nonnative male speakers, and that emphatic phonemes yield a significant decrease in performance when they are uttered by both male and female nonnative speakers.

1. INTRODUCTION

Pronunciation variability is by far the most critical issue for Arabic automatic speech recognition (AASR). This is mainly due to the large number of nonnative accents and to the fact that nonnative speech data available for training are generally insufficient. Hence the modeling of separate accents remains difficult and inaccurate. In addition, the Arabic language is characterized by an extreme dialectal variation and nonstandardized speech representations, since it is usually written without short vowels and other diacritics, and thus has incomplete phonetic information [1].

During the past few years, there have been research initiatives carried out on analyzing speech from native and nonnative speakers' points of view. Byrne et al. [2] worked on analyzing nonnative English speakers by collecting a corpus of conversational English speech from nonnative English speakers. They used an HTK-based speech recognition system. Their corpus contained both read and conversational speech recordings. They concluded that it is hard to recognize nonnative English speakers compared to native ones especially with regard to conversational type. Another study was carried out by Livescu [3]. His work concentrated on analyzing and modeling nonnative speech for automatic speech recognition. He examined—among other tasks—the problem of nonnative speech in a speaker independent, large-vocabulary, spontaneous speech recognition system for American English with native training data. He showed that the interpolated native and nonnative models reduce the word error rate on a nonnative test set by 8.1% relative to his baseline recognizer using models trained on pooled native and nonnative data. He also investigated many issues

in language model (LM) differences in native and nonnative speakers. To improve the performance of the speech recognition system for nonnative speakers, Bartkova and Jouvet [4] propose an approach based on multiple models. They considered French as the native language. In their study, they included English, Spanish, Italian, Portuguese, Turkish, and Arabic nonnative groups. This approach required a huge amount of training data. Compared with research on other languages, only a very limited number of research initiatives have been carried out on Arabic language.

The aim of this paper is to investigate the effect of foreign accents for both male and female speakers on the performance of automatic speech recognition of Arabic. In this way, we can figure out the effects of these variations on the overall HMM-based system accuracy using a language model (LM), and on the individual phoneme accuracy of an HMM-based system, which does not use an LM.

This paper is organized as follows. Section 2 summarizes the main characteristics of the Arabic language. Section 3 describes the data and the baseline systems used in this study. Section 4 presents and discusses the obtained results. Section 5 concludes and indicates the perspective of this work.

2. ARABIC LANGUAGE CHARACTERISTICS

Arabic is a Semitic language, and it is one of the oldest languages in the world today. It is the fifth widely used language nowadays. Arabic is the first language in the Arab world, used in Saudi Arabia, Jordan, Oman, Yemen, Egypt, Syria, Lebanon, and many more countries. The Arabic alphabet is used in several languages, such as Persian, Urdu, and Malay [5]. Research on the Arabic language has mainly concentrated on modern standard Arabic, which is used throughout the media, courtrooms, and academic institutions in Arab countries. Previous work on developing ASR was dedicated to dialectal and colloquial Arabic within the 1997 NIST benchmark evaluations, and more recently on the recognition of conversational and dialectal speech, as it is reported in [1].

2.1. Phonetic features

The standard Arabic language has 34 phonemes, of which six are basic vowels, and 28 are consonants. The Arabic language has fewer vowels than the English language. It has three long and three short vowels, while American English has at least twelve vowels. Standard Arabic is distinct from Indo-European languages because of its consonantal nature. The allowed syllable structures in Arabic are CV, CVC, and CVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant [6]. From an articulatory point of view, it is characterized by the realization of some sounds in the rear part of the vocal tract: glottal and pharyngeal consonants. Arabic sounds can be divided into macroclasses such as stop consonants, voiceless fricatives, voiced fricatives, nasal consonants, liquid consonants, and vowels. The originality of the Arabic phonetics is mainly based on the relevance of lengthening in the vocalic system and on the presence of emphatic and geminated consonants. These particular features play a fundamental role in the nominal and verbal morphological development. Pharyngeal and emphatic phonemes exist only in Semitic languages like Hebrew, Persian, and Urdu [6, 7].

Emphatic consonants are achieved in the rear part of the oral cavity. During their production, the root of the tongue is carried against the pharynx. There are four emphatic consonants in the Arabic language: two plosives: /t̲/, /d̲/ and two fricatives: /ð /, /ṣ/. In the example of the two words /naṣaba/ (imputed) and /nasaba/ (erected), an emphatic versus nonemphatic opposition is observed on /s/ [7].

The gemination is a particular feature, which compensates for the paucity of the Arabic vocalic system. The geminated consonant arises by sustaining the plosive closure. In the example of the words /fa ala/ (he failed) and /fa :ala/ (he thwarts), the opposition resides in the gemination of / / fricative. Through this example, we measure the importance and the difficulty of performing this feature detection.

The vocalic system contains two phonological quantities for each tone. For each short vowel /a/, /i/, and /u/, there is, respectively, the associated long vowel /a:/, /i:/, and /u:/. In Arabic, this temporal opposition is fundamental. For example, the two words /3amal/ "camel" and /3ama:l/ "beauty", have the length of the final vowel as the only difference [8].

2.2. Morphological complexity

The development of accurate AASR systems is faced with two major issues. The first problem is related to diacritization. Arabic texts are almost never fully diacritized: it means that the short strokes placed above or below the consonant, indicating the vowel following this consonant, are usually absent. This limits the availability of training material. The lack of this information leads to many similar word forms, and then decreases predictability in the language model. The second problem is related to the morphological complexity since Arabic has a rich potential of word forms which increases the out-vocabulary rate [8, 9].

3. DATA AND BASELINE SYSTEMS

3.1. The WestPoint corpus

The WestPoint Arabic corpus, provided by LDC [10], is used in our experiments. It consists of collections of four main Arabic scripts. The first one is Collection Script 1, which contains 155 sentences, used by all 74 native Arabic speakers. Script 1 has a total of 1152 tokens and 724 types. The second one is Collection Script 2, which contains 40 sentences, used by 23 of the nonnative speakers. Script 2 has a total of 150 tokens and 124 types. The third one is Collection Script 3, which contains 41 sentences, used by 4 of the nonnative speakers. It has a total of 138 tokens and 84 types. Finally, there is Collection Script 4, which contains 22 sentences, used by 9 of the nonnative 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 1,131 Arabic words. All scripts

Table 1: Arabic phoneme list used throughout our experiments.

LDC phoneme	Description	IPA symbol
C	voiced pharyngeal fricative	
D	velarized voiced alveolar stop	
G	voiced velar fricative	ɣ
H	voiceless pharyngeal fricative	
Q	voiceless glottal stop	
S	velarized voiceless alveolar fricative	ṣ
T	velarized voiceless alveolar stop	ṭ
TH	velarized voiced interdental fricative	ð
Z	voiced interdental fricative	ð
ae	low front vowel	aa
ah	low back vowel	a
aw	back upgliding diphthong	a-w
ay	front upgliding diphthong	a-i
b	bilabial voiced stop	b
d	voiced alveolar stop	d
ey	upper mid front vowel	a-y
f	voiceless labiodental fricative	f
g	voiced velar stop	g
h	voiceless glottal fricative	h
ih	high front lax vowel	I
iy	high front tense vowel	ii
j	voiced palato-alveolar fricative	3
k	voiceless velar stop	k
l	voiced alveolar lateral	l
m	voiced bilabial nasal	m
n	voiced alveolar nasal	n
q	voiceless uvular stop	q
r	voiced alveolar flap	r
s	voiceless alveolar fricative	s
sh	voiceless palato-alveolar fricative	ʃ
t	voiceless alveolar stop	t
th	voiceless interdental fricative	θ
uw	high back rounded vowel	o
w	voiced bilabial approximant	w
x	voiceless velar fricative	x
y	voiced palatal approximant	j
z	voiced alveolar fricative	z

Table 2: WestPoint corpus statistical summary.

Number of speakers			
	male	female	total
native	41	34	75
non-native	25	10	35
totals	66	44	110
Hours of data			
	male	female	total
native	6	4.4	10.4
non-native	0.74	0.28	1.02
totals	6.74	4.68	11.42
Megabyte of data			
	male	female	total
native	913	663	1576
non-native	111	42.4	153.4
totals	1024	705.4	1729.4
Number if speech files			
	male	female	total
native	4107	3163	7270
non-native	883	363	1246
totals	4990	3526	8516

Table 3: Experimental conditions summary.

Parameter	Value
Sampling rate	22.05 KHz, 16 bits
Database	LDC2002S02 (WestPoint)
Speakers	44 Female + 66 Male
Features	MFCCs with first derivative
Preemphased	$1–0.95\,z^{-1}$
Window type and size	Hamming, 256
Window step size	64
Order	12

were written with MSA as the target language and were diacritized. Table 1 shows the Arabic phonemes and their symbols used with LDC along with international phonetic alphabet (IPA). A summary of the statistical numbers of this database is given in Table 2. As we can see from this table, the amount of data provided by the Arabic native speakers is significantly bigger than that of the data provided by the Arabic nonnative speakers.

3.2. The parameterization

The parameters of the system are 22.05 KHz sampling rate with 16 bit sample resolution, 25-milliseconds Hamming window duration with a step size of 10 milliseconds, MFCCs with 22 as the length of cepstral liftering, 26 as the number of filter bank channels, 12 as the number of MFCC coefficients, and 0.95 as the pre-emphasis coefficient. Table 3 shows the details of the system parameters.

3.3. The recognizer

The Hidden Markov Model Toolkit (HTK) [11] is used for designing and testing the speech recognition systems throughout all experiments. The baseline system was initially designed as a phoneme level recognizer with three active states, one Gaussian per state, continuous, left-to-right, and no skip HMM models. The system was designed by considering all 37 MSA phones as given by the LDC West-Point catalog. The WestPoint corpus has three phonemes more than the number of MSA phonemes mentioned in most linguistic literature [6, 7, 9]. WestPoint added three more phonemes, namely, /g/ "voiced velar stop", /aw/ "back

upgliding diphthong", and /ey/ "upper mid front vowel". In fact, the phoneme /g/ does not exist in MSA at all, but we think that the WestPoint corpus used it because some native and nonnative speakers are using it popularly in some MSA words. We can confirm this fact by hearing some WestPoint audio files. On the other hand, the extra vowel and diphthong were used because of variations in pronunciations of speakers influenced by English and other Latin languages. This type of phoneme exists in these languages but not in MSA. For our study, we finally decided to stick with WestPoint phonemes and transcriptions without any modification. We believe that this decision will facilitate the comparison with systems of other research efforts that are using the same corpus. Since most words consisted of more than two phonemes, context-dependent triphone models were created from monophone models. The training phase consists of re-estimating HMM models by using the Baum-Welch algorithm after aligning and tying the models by using the decision tree method [12]. Phoneme-based models are good at capturing phonetic details. Also context-dependent phoneme models can be used to characterize formant transition information, which is very important in the discrimination of confusable speech units.

3.4. The language model

The performance of any recognition system depends on many factors, but the size and the perplexity of the vocabulary are among the most critical ones. In this system, the size of vocabulary is relatively high since it contains more than one thousand different words. Their perplexity is very high due to the existence of many acoustically similar phonemes in Arabic.

A language model is essential for effective speech recognition. Typically, the LM will restrict the allowed sequences of words in an utterance. It can be expressed by the formula giving the a priori probability, $P(W)$:

$$P(W) = p(w_1, \ldots w_m) = p(w_1) \prod_{i=2}^{m} p\left(w_i \mid \underbrace{w_{i-n+1}, \ldots w_{i-1}}_{n-1}\right), \tag{1}$$

where $W = w_1, \ldots w_m$ is the sequence of words. In the n-gram approach described by (1), n is typically restricted to $n = 2$ (bigram) or $n = 3$ (trigram).

The language model used in our experiments is a bigram, which mainly depends on the statistical numbers that were generated from the phonetic transcriptions of all words of both the training and the test directories of the WestPoint corpus. All input transcriptions (labels) are fed to a set of unique integers in the range 1 to L, where L is the number of distinct labels. For each adjacent pair of labels i and j, the total number of occurrences $O(i, j)$ is counted. For a given label i, the total number of occurrences is given by

$$O(i) = \sum_{j=1}^{L} O(i, j). \tag{2}$$

For both word and phonetic matrix bigrams, the bigram probability $p(i, j)$ is given by

$$p(i, j) = \begin{cases} \alpha \dfrac{O(i, j)}{O(i)} & \text{if } O(i) > 0, \\ \dfrac{1}{L} & \text{if } O(i) = 0, \\ \beta & \text{otherwise,} \end{cases} \tag{3}$$

where β is a floor probability and α is chosen to ensure that

$$\sum_{j=1}^{L} p(i, j) = 1. \tag{4}$$

For back off bigrams, the unigram probablities $p(i)$ are given by

$$p(i) = \begin{cases} \dfrac{O(i)}{O} & \text{if } O(i) > \gamma, \\ \dfrac{\gamma}{O} & \text{otherwise,} \end{cases} \tag{5}$$

where γ is unigram floor count and O is determined as follows:

$$O = \sum_{j=1}^{L} \max\left[O(i), \gamma\right]. \tag{6}$$

The backed off bigram probabilities are given by

$$p(i, j) = \begin{cases} \dfrac{(O(i, j) - D)}{O(i)} & \text{if } O(i, j) > \theta, \\ b(i)p(j) & \text{otherwise,} \end{cases} \tag{7}$$

where D is a discount and θ is a bigram count threshold. The discount D is fixed at 0.5. The back off weight $b(i)$ is calculated to ensure that

$$\sum_{j=1}^{L} p(i, j) = 1. \tag{8}$$

The bigram probability is used because at least one bigram has been observed in the training data; otherwise the transition probability is calculated from the unigram count. These statistics are generated by using HLStats function, which is a tool of the HTK toolkit. This function computes the occurrences of all labels in the system and then it generates the back off bigram probabilities based on the phoneme-based dictionary of the corpus. This file counts the probability of the occurrences of every consecutive pairs of labels in all labelled words of our dictionary. A second function of HTK toolkit, HBuild, uses the back off probabilities file as an input and generates the bigram language model. The dictionary used in our application includes all (without any exception) words that were used in WestPoint corpus.

Table 4: System overall performance with different configurations with respect to native origin of speakers and using the Arabic LDC-WestPoint corpus.

	N/N Exp. 1.a	N/NN Exp. 2.a	NN/N Exp. 3.a	NN/NN Exp. 4.a	M/M Exp. 5.a	N/N Exp. 1.b	N/NN Exp. 2.b	NN/N Exp. 3.b	NN/NN Exp. 4.b	M/M Exp. 5.b
Corr (%)	52,01%	46,44%	43,03%	**54,54%**	50,32%	**99,05%**	89,46%	93,98%	96,78%	98,85%
Del (%)	2,33%	2,24%	2,43%	2,74%	2,37%	0,00%	0,00%	0,01%	0,00%	0,00%
Sub (%)	45,66%	51,32%	54,54%	4,27%	47,31%	0,01%	10,22%	5,08%	2,89%	0,01%
	Without bigram-based language model					With bigram-based language model				

4. EXPERIMENTS, RESULTS, AND DISCUSSION

Nine sets of experiments: Exp. 1, Exp. 2, Exp. 3, ..., Exp. 9 have been carried out. In each experiment we examined two outcomes from the system. The first outcome concerns phonemic recognition without using any LM. It is referred to by the *a* subscript. The language dictionary used in the system is a simple phoneme-to-phoneme mapping. The second outcome, referred to by the subscript *b*, consists of the system accuracy when the LM is incorporated. It uses a dictionary, mapping every word in the database including its corresponding phoneme transcription. In all the nine experiments, the difference is the type of training and testing database sets depending on a speakers' native language. As specified by the WestPoint database, if the speaker is a nonnative Arabic speaker, this means that he or she is an English native speaker.

4.1. Experimental setup

In the first experiment, Exp. 1, native Arabic speakers are involved in both the training and the testing. The WestPoint corpus is divided in such a way that 61% of the corpus is used for the training and 39% for the test regardless of either gender or scripts. In the second experiment, Exp. 2, native speakers are used for the training and only nonnative speakers are involved in the test phase. Then, the training uses 85% of the corpus (native speakers), while 15% of the corpus composed of nonnative speakers is used for the test. In Exp. 3, all nonnative Arabic speakers were used for the training and all Arabic native speakers for the test. In the fourth experiment, Exp. 4, only nonnative Arabic speakers were used in both training and testing systems. The nonnative speakers' part of the corpus is divided to obtain 67% for the training, and 33% for the test. In the fifth experiment, Exp. 5, both native and nonnative Arabic utterances are pooled to constitute the training data (69% of the corpus) and testing data (31% of the corpus). Experiments Exp. 6 to Exp. 9 are set up by varying the gender of the speakers in training and testing data. In Exp. 6, native male speakers are used for the training and only nonnative male speakers are involved in the test. Thus the training uses 82% of the corpus (male speakers only), while 18% of the corpus composed of nonnative speakers is used for the test. In Exp. 7, native female speakers are used for the training and only nonnative female speakers are involved in the test. Hence the training uses 90% of the corpus (female speakers only), while 10% of the corpus composed of nonnative speakers is used for the test. In Exp. 8, nonnative male speakers are used for the training and only native male speakers are involved in the test. In this case, the training uses 18% of the corpus (male speakers), while 82% of the corpus composed of native male speakers is used for the test. Finally, in Exp. 9, nonnative female speakers are used for the training and only native female speakers are involved in the test. Thus the training uses 10% of the corpus (female speakers), while 90% of the corpus composed of nonnative speakers is used for the test. The bigram language model was always derived from the total of prompts provided by WestPoint. This means that we use the same language model for all experiments.

Table 5: System overall performance with different gender and native language configurations.

Train/Test	All phoneme level without any language model	Word level with a language model
Male N/NN	Exp. 6a: 46.56%	Exp. 6b: 89.07%
Female N/NN	Exp. 7a: 51.42%	Exp. 7b: 83.62%
Male NN/N	Exp. 8a: 47.86%	Exp. 8b: 95.89%
Female NN/N	Exp. 9a: 50.17%	Exp. 9b: 81.34%

4.2. Effect of the language model

Tables 4 and 5 present the overall system accuracies of the nine experiments in both word level (using LM) and phoneme level (without using any LM) by considering the same probability of any two sequences of phonemes. A set of experiments is carried out by incorporating a language model with the triphone HMM-based system (referred to by *b* index in both of Tables 4 and 5). If the considered units for the accuracy of AASR are words, the overall performance of the system is increased by around 50%, as shown in Table 4 (experiments Exp. 1(b) through Exp. 5(b)). The best accuracy is obtained in Exp. 1(b) where both the training and testing data set use native speakers. In the mixed mode (Exp. 5(b)), the accuracy reaches 98.85%. As for the phonetic level, the accuracy drops down if the training and testing data sets are not identical with respect to the native origin of speakers in both sets as observed in Exp. 2(b) and Exp. 3(b). Note that when the LM is introduced, and if the training and testing sets are different with respect to native origin of speakers, better accuracy (more than 4%) is obtained

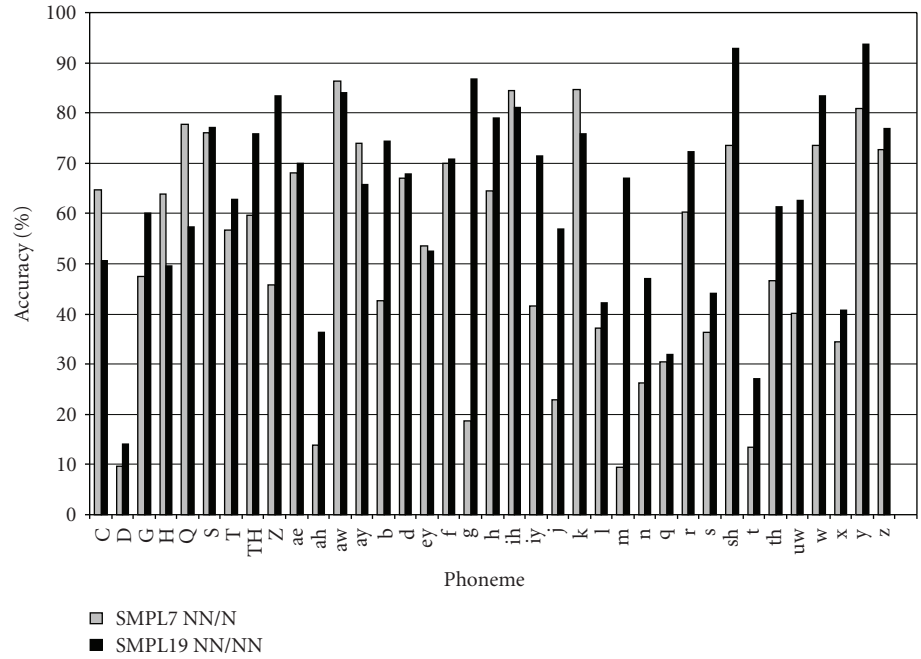

Figure 1: Phoneme accuracies for experiments Exp. 3(a) and Exp. 4(b).

when nonnative speakers perform the training and the native ones perform the test (i.e., in Exp. 3(b) compared to Exp. 2(b)). At the phoneme level of accuracy, we observe that the hierarchy is completely inverted. This leads us to conclude that the introduction of LM in the AASR masks numerous pronunciation errors due to foreign accents. These errors are investigated in more detail in the following sections.

4.3. Effect of the native origin of speakers

If we consider the phoneme recognition performance without using any LM, as shown in Table 4, we notice that the system gives its best accuracy when it is trained and tested by nonnative speakers. The poorest overall accuracy is obtained when the system is trained on nonnative speakers and tested on native speakers. In other words, the nonnative trained system gave the best accuracy and the worst accuracy if tested by nonnative and native speakers, respectively. By investigating the detailed results that are related to the accuracy of each phoneme, we found that some phonemes give lower accuracy if tested with native speakers instead of nonnative ones. In fact, the best phoneme recognition rate is reached when nonnative speakers are involved in the training and in the test phases. This result can be explained by the fact that nonnative speakers tend to make efforts in order to be more consistent with the standard pronunciation. When the training and testing data sets in each experiment are identical with respect to the native origin of the speakers, the accuracies are higher compared to the cases where training and test sets are different with respect to the native origin of the speakers. If the training and testing data sets are mixed (regardless of the native origin of speakers), the accuracy decreases by almost 2% and 4% compared to the results obtained in Exp. 1(a), and Exp. 4(a), respectively. As expected, the accuracy of an AASR system is negatively influenced by changing the mother tongues in either the training or testing data sets.

4.4. Effect of the speakers' gender

The gender of the speaker is one of the influential sources of speech variability. In the early days of speech recognition, gender was not considered as a major issue. The progress made last decade led to high performance transcription systems that permit one to consider the question whether ASR systems behave differently on male and female speech. An interesting study carried out by Adda-Decker and Lamel [13] reveled that for both French and English languages, female speakers had better average recognition results than males. In our experiments, and by considering the gender of speakers, as it can be inferred from Table 5 (without a LM, i.e., experiments that numbered by subscript b), we can notice that female speakers give better system overall accuracy. This difference is more than 2% in case where nonnative speakers are involved in the training and the native ones in the test. On the other hand, the improvement of using female speakers is almost 5% when native speakers are used in the training and nonnative speakers in the test. By incorporating a LM and by considering the word level (i.e., in experiments Exp. 6(b) through Exp. 9(b), we see that the argument is inverted. In other words, the LM improved the accuracy of male speakers in a much better way than in the case of female speakers.

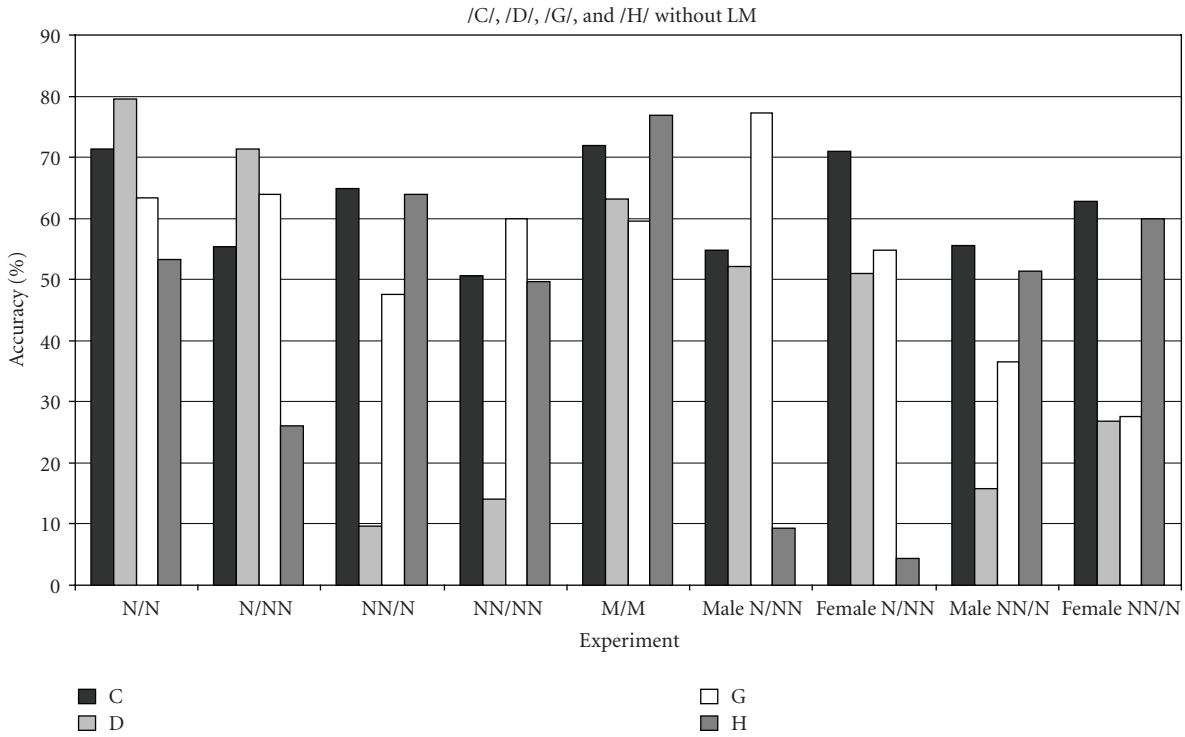

Figure 2: Distinct Arabic phonemes accuracy in all experiments—Group 1.

4.5. Performance at the phonetic level

If we train the system by using nonnative speakers and test it separately, firstly by nonnative and secondly by native speakers (Exp. 4(a) and Exp. 3(a)), we noticed that phonemes /G/, /TH/, /Z/, /ah/, /b/, /g/, /h/, /iy/, /j/, /m/, /n/, /r/, /sh/, /t/, th/, /uw/, /w/, and /y/ decrease their accuracies by more than 10%. On the other hand, phonemes /C/, /H/, /ay/, and /k/ got better accuracies. As can be figured out from Figure 1, this dramatic drop (around or larger than 10%) of phoneme accuracies for 18 phonemes leads one to investigate more deeply the phonemes that are at the root of the gap between native speakers "and nonnative speakers" performance. It is worthy to note that nonnative speakers have significant problems with the pronunciation of specific Arabic phonemes such as the emphatic /D/ and the voiceless stop consonant /t/. The emphatic /D/ is the symbol of the Arabic language since it may only exist in Arabic, and it is very difficult to pronounce for a non-Arabic speaker. For the phoneme /t/, there was more than a 10% difference between its accuracy in Exp. 3(a) and Exp. 4(a) as can be noted in Figure 1. This implied that native and nonnative Arabic speakers, indeed, pronounce this phoneme in two different ways. We believe that it is due to the difference in the place of articulation (here the position of tongue dip when the /t/ is uttered) of English /t/ and Arabic /t/. This difference is noticeable by hearing Arabic and English speech in the case of this phoneme pronunciation.

Figures 2 and 3 give special attention to specific Arabic phonemes that cannot be found in Latin languages, especially English. These figures plot the accuracy of every phoneme of this set with respect to each set of the nine experiments. Phonemes /C/, /D/, /G/, and /H/ are shown in Figure 2, while phonemes /S/, /T/, /Z/, and /x/ are shown in Figure 3. By analyzing these two figures, we can conclude that, globally, the individual accuracies of these phonemes are improved when native and nonnative speakers' utterances are pooled in the training and testing data sets.

The Arabic phoneme /H/ is a pharyngeal, fricative, unvoiced, and nonemphatic sound. The /H/ phoneme sharply falls in accuracy whenever nonnative speakers are involved in training and/or testing data. The accuracy of this phoneme is less than 10% in experiments Exp. 6(a) (Male N/NN) and Exp. 7(a) (Female N/NN). On the other hand, this phoneme gives better performance in the other experiments.

The phoneme /G/, which is an alveo-dental, stop, voiced, and emphatic sound, is similar to the /H/ phoneme. It sharply drops in accuracy whenever nonnative speakers are involved in training and/or testing data. To cite as examples, the accuracy of this phoneme is less than 20% in experiments Exp. 3(a) (NN/N) and Exp. 4(a) (NN/NN). It gives a better performance in other experiments.

5. CONCLUSION

In this paper, we have presented the results obtained by an HMM-based speaker independent, large-vocabulary speech recognition system for modern standard Arabic with a focus on the problem of foreign accents. We analyzed the

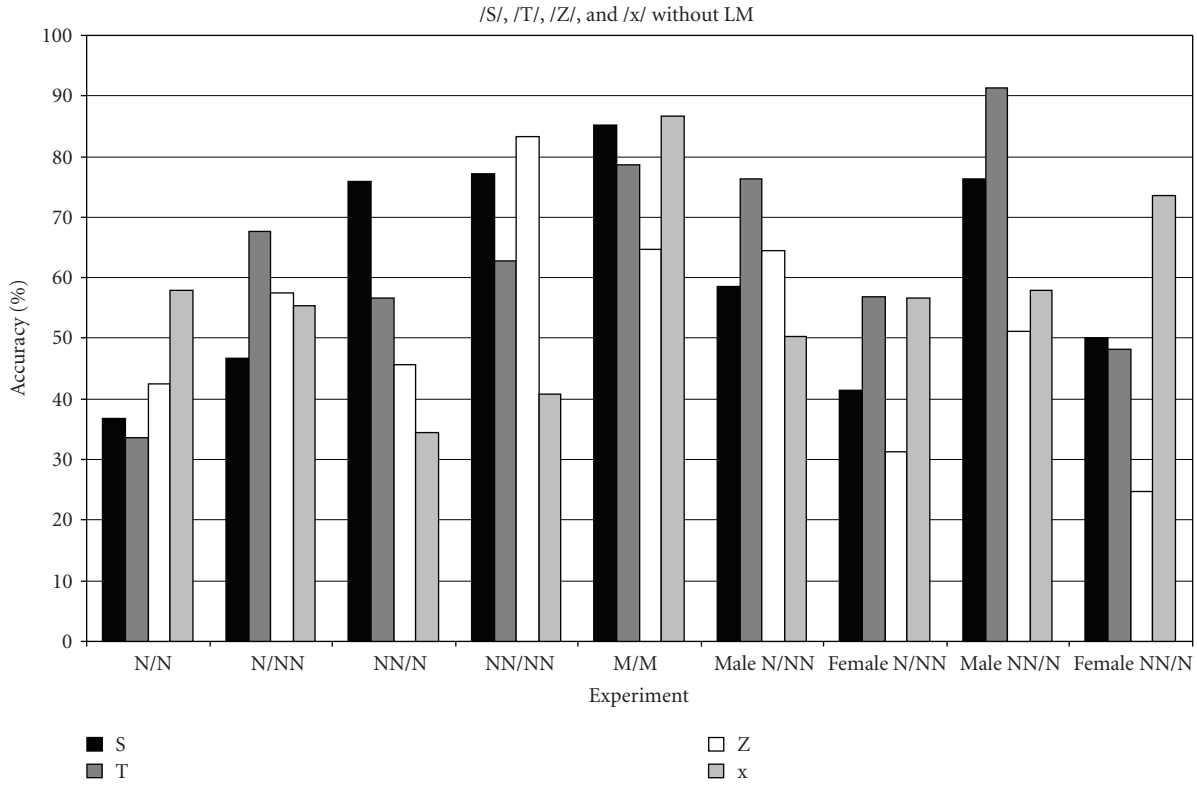

Figure 3: Distinct Arabic phonemes accuracy in all experiments—Group 2.

performance of AASR at phonetic and word levels. We have confirmed, through our experiments, that the accuracy of the AASR system is negatively influenced by changing the mother tongues in either the training or testing data sets. However, the best phoneme recognition rate is reached when nonnative speakers are involved in both the training and the test phases, which is far from being predictable. The obtained results show that at the phonetic level, the female nonnative speakers perform better than nonnative male speakers. These results confirm that as it is observed for English and French languages [13], the pronunciation of nonnative Arabic female speakers tends to be more consistent with the standard pronunciation than that of the nonnative male speakers. However, the bigram-based language model improved the accuracy of nonnative male speakers in a much better way than that of the case of female speakers. In addition, we have noticed that nonnative speakers have difficulty in pronouncing the /D/ emphatic consonant. We must note here that the /D/ is a unique phoneme that exists only in Arabic. It is the reason why the Arabic language is commonly known as "the /D/ language" by the Arab community. It is worthy to note that nonnative speakers have significant problems with the pronunciation of the voiceless stop consonant /t/. There was more than a 10% difference between native and nonnative accuracies. We confirmed by hearing all WestPoint Corpus utterances that contain this phoneme, that it is due to the difference in the place of articulation, that is, in the position of the tongue dip when the /t/ is uttered by native and nonnative speakers. We will continue this research work by investigating the best way to adapt the AASR system to foreign accents by introducing the phonetic knowledge acquired from the common errors of nonnative speakers.

REFERENCES

[1] K. Kirchhoff, J. Bilmes, S. Das, et al., "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 344–347, Hong Kong, April 2003.

[2] W. Byrne, E. Knodt, S. Khudanpur, and J. Bernstein, "Is automatic speech recognition ready for non-native speech? A data collection effort and initial experiments in modeling conversational Hispanic English," in *Proceedings of the ESCA Conference on Speech Technology in Language Learning (STILL '98)*, pp. 37–40, Marholmen, Sweeden, May 1998.

[3] K. Livescu, "Analysis and modeling of non-native speech for automatic speech recognition," M.S. thesis, MIT, Cambridge, Mass, USA, 1999.

[4] K. Bartkova and D. Jouvet, "Multiple models for improved speech recognition for non-native speakers," in *Proceedings of the 9th Conference of Speech and Computer (SPECOM '04)*, St. Petersburg, Russia, September 2004.

[5] M. Al-Zabibi, "An acoustic-phonetic approach in automatic Arabic speech recognition," Ph.D. thesis, Loughborough University of Technology, Leics, UK, 1990.

[6] M. Alkhouli, *Linguistic Phonetics*, Daar Alfalah, Swaileh, Jordan, 1990.

[7] M. Elshafei, "Toward an Arabic text-to-speech system," *The Arabian Journal for Science and Engineering*, vol. 16, no. 4, pp. 565–583, 1991.

[8] Y. M. El-Imam, "An unrestricted vocabulary Arabic speech synthesis system," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1829–1845, 1989.

[9] A. Omar, *Studying Linguistic Phonetics*, Aalam Alkutob, Cairo, Egypt, 1991.

[10] Linguistic Data Consortium (LDC) Catalog Number LDC2002S02, 2002, http://www.ldc.upenn.edu/.

[11] S. Young, G. Evermann, M. Gales, et al., "The HTK Book (for HTK Version. 3.3)," Cambridge University Engineering Department, 2005, http://htk.eng.cam.ac.uk/.

[12] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*, Marcel Dekker, New York, NY, USA, 2003.

[13] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Proceedings of the InterSpeech Conference (Interspeech '05)*, pp. 2205–2208, Lisbon, Portugal, September 2005.