J. King Saud Univ., Vol. 21, *Comp. & Info. Sci.*, pp. 11-25, Riyadh (2009/1430H.)

Investigating Emphatic Consonants in Foreign Accented Arabic

Yousef Ajami Alotaibi[1], Sid-Ahmed Selouani[2] and Wladyslaw Cichocki[3]
[1]*Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Saudi Arabia*
[2]*LARIHS Laboratory, Université de Moncton, Campus de Shippagan, Canada*
[3]*Department of French, University of New Brunswick, Canada*

(Received 21/11/2007; accepted for publication 25/01/2009)

Keywords: Arabic, Emphatic consonants, Speech recognition, Foreign accents, Spectrograms, HMMs.

Abstract. This paper investigates the four emphatic consonants of Arabic from the point of view of automatic speech recognition. Comparisons of the recognition error rates for these phonemes and for their non-emphatic counterparts are analyzed in five experiments that involve different combinations of native and non-native Arabic speakers. In addition, the target consonants are described in time-frequency domain analyses. All experiments used the Hidden Markov Model toolkit (HTK) and the Language Data Consortium (LDC) WestPoint Modern Standard Arabic (MSA) database. Results confirm that emphatic consonants are a major source of difficulty for ASR. While the recognition rate for certain emphatic consonants such as /D/ can drop below 15% when uttered by non-native speakers, there are advantages to including non-native speakers in ASR. Regional differences in the pronunciation of MSA by native Arabic speakers require the attention of Arabic ASR research.

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, and pharyngeal consonants, and vowel duration (Selouani and Caelen, 1998; Economic and Social Commission for Western Asia – 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 four Arabic emphatic 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 it explains related topics that will give readers an overview of some of the difficulties that Arabic ASR faces, including those with emphatic consonants.

1.1. Arabic language

Arabic 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 (Kirchhoff *et al.*, 2002). 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 (Al-Zabibi, 1990; Alkhouli, 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).

Arabic is written in script and from right to left. The alphabet consists of 29 letters, 26 of which represent consonants. The remaining 3 letters represents the long vowels of Arabic (the phonemes /i:/, /a:/, /u:/) and, where applicable, the corresponding semivowels (the phonemes /y/ and /w/). Each letter can appear in up to four different shapes, depending on whether it occurs at the beginning, in the middle, or at end of a word, or in isolation (Omar, 1991). A distinguishing feature of the Arabic writing system is that short vowels and consonant doubling are not represented by the letters of the alphabet. Instead, they are marked by so-called diacritics, short strokes placed either above or below the preceding consonant (Elshafei, 1991). However, Arabic texts are almost never fully diacritized and are thus potentially unsuitable for automatic digital speech processing such as speech recognition and synthesis (Kirchhoff *et al.*, 2003). 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. In Table 1 and throughout this paper, we use the symbols of Language Data Consortium (LDC) WestPoint Modern MSA database phoneme set rather than those of the International Phonetic Alphabet (IPA).

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 34 phonemes consisting of three short vowels (/i/, /a/, /u/), three long vowels (/i:/, /a:/, /u:/ which are the counterparts of the short vowels), and 28 consonants (Alghamdi, 2004).

Arabic has noticeably fewer vowels than English. While some varieties of American English have at least 12 vowels, Arabic has three long and three short vowels (Deller *et al.*, 1993). In addition, vowel lengthening in Arabic is phonemic. Some Arabic dialects may have additional or fewer consonant phonemes. For example, Egyptian Arabic dialect does not use the phonemes /TH/ and /th/, and it replaces phoneme /j/ with phoneme /g/ (Kirchhoff *et al.*, 2002). Arabic phonemes contain two distinctive classes that are named pharyngeal and emphatic phonemes. These two classes are found in Semitic languages like Hebrew and Arabic (Alkhouli, 1990; Elshafei, 1991).

The co-articulation effect caused by emphatic phonemes can affect adjacent phonemes especially vowels. The emphatic consonants induce a considerable backing (i.e., relatively moving the tongue back during articulation) gesture in neighboring segments, which occurs primarily for adjacent vowels. This effect may spread over entire syllables and beyond syllable boundaries (El-Imam, 2001). It is not easy to determine the extent of the co-articulation effect of the emphatic and pharyngeal phonemes on their neighboring consonants and vowels (Laufer and Baer, 1988; Ouni *et al.*, 2005; Watson, 1999).

The syllable types that are allowed in the Arabic language are CV, CVC, and CVCC, where V indicates a (long or short) vowel and C indicates a consonant; the vowel in the third type of mentioned Arabic syllables can be short only (Alghamdi, 2001). 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. 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).

Arabic has a rich and productive morphology, which leads to a large number of potential word forms. This increases the out-of-vocabulary rate and prevents the robust estimation of LM probabilities (Kirchhoff *et al.*, 2003). Much of the complexity of the Arabic language is found at the morphological level. Arabic has two genders (masculine, feminine), three numbers (singular, dual, plural), three cases (subject case, object case, prepositional object case) and two morphologically marked tenses. There is noun-adjective agreement for number and gender, and there is subject-verb agreement (Omar, 1991; (Kirchhoff *et al.*, 2002).

Table 1. MSA Arabic consonants (Alghamdi, 2001)

			Bilabial	Labio-dental	Inter-dental	Alveo-dental	Alveolar	Palatal	Velar	Uvular	Pharyngeal	Glottal
Stop	Voiced	Emphatic				ض /D/						
		Non-emphatic	ب /b/			د /d/						
	Unvoiced	Emphatic				ط /T/						
		Non-emphatic				ت /t/			ك /k/	ق /q/		ء /Q/
Fricative	Voiced	Emphatic			ظ /Z/							
		Non-emphatic			ذ /TH/	ز /z/				غ /G/	ع /C/	
	Unvoiced	Emphatic				ص /S/						
		Non-emphatic		ف /f/	ث /th/	س /s/		ش /sh/		خ /x/	ح /H/	هـ /h/
Nasal	Voiced	Non-emphatic	م /m/				ن /n/					
Lateral/Trill	Voiced	Non-emphatic					ر ل /l/r/					
		Emphatic					ل /L/					
Semivowels	Voiced	Non-emphatic	و /w/					ي /y/				
Affricate	Voiced	Non-emphatic						ج /j/				

1.3. Emphatic consonants in Arabic

There are four emphatic consonants in Arabic that are of interest here: two plosives, /D/ and /T/, and two fricatives, /S/ and /Z/ (Selouani and Caelen, 1998; Al-Muhtaseb *et al.*, 2000; Ouni *et al.*, 2005). /D/ is a voiced empathic plosive with an alveo-dental point of articulation. As this phoneme 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. Moreover, this name was given to Arabic depending on the classical Arabic version of /D/ phoneme which is an emphatic lateral fricative, but not plosive as given by MSA version. /T/ is an unvoiced emphatic plosive with an alveo-dental point of articulation. /S/ is an unvoiced emphatic fricative with an alveo-

dental point of articulation. Finally, /Z/ is a voiced empathic fricative with an inter-dental point of articulation (Alkhouli, 1990). Table 2 shows the four emphatic Arabic sounds and their non-emphatic counterparts. The uvular fricative /G/ is not studied in this investigation.

There is a noteworthy exception regarding the phonemes /r/ and /l/. These may become emphatic in certain limited cases in Classical Arabic and in MSA (Alghamdi, 2004). The phoneme /l/ is an emphatic sound in the word God "ALLaah", pronounced in Arabic as "*waLLah*", but in all other cases /l/ is a non-emphatic sound. In our example, if we change the phoneme /l/ to a non-empathic sound, the meaning of that word is "he appointed him" and not "God". The emphatic-ness may also affect the phoneme /r/ in similar way as the /l/ phoneme.

Table 2. Arabic empathic sounds and their non-emphatic counterparts

Arabic Alphabet Carrier	LDC Symbol	IPA Symbol	Non-empahtic Counterparts
Dhaad ض	D	$d^{?}$	/d/ Daal
Saad ص	S	$s^{?}$	Voiced: /z/ (Zain); Unvoiced: /s/ (Seen)
T_aa ط	T	$t^{?}$	Voiced: /d/ (Daal); Unvoiced: /t/ (Taa)
Dhaa ظ	Z	$D^{?}$	/TH/ (Thaal)

Several factors can affect the pronunciation of phonemes including their position in the syllable, either initial or final, or in suffixes. The pronunciation of consonants may also be influenced by co-articulation with phonemes in the same syllable. Among these effects are pharyngealization and nasalization. Arabic vowels are affected as well by the adjacent phonemes. Accordingly, each Arabic vowel has at least three allophones: a normal, an emphatic, and a nasalized allophone (Alghamdi, 2004). Some dialects show labialization of vowels in the environment of emphatics (Watson, 1996).

1.4. Emphatic sounds in 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.

Selouani 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 /D/ 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.

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 four emphatic consonants—/S/, /D/, /T/, and /Z/—and their non-empathic counterparts—the /s/, /d/, /t/ and /TH/ consonants. The study focuses on the effect of native and non-native speakers in both training and testing data. The accuracy with respect to all eight segments is investigated in detail.

2.1. ASR technique

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 process, 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 (Rabiner, 1989; Juang and Rabiner, 1991). 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

We used the WestPoint Arabic Speech Corpus, provided by LDC (Linguistic Data Consortium, 2002), in our experiments. 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 1,152 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 contains 41 sentences used by 4 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 9 of the non-native speakers, all of them third-year Arabic learners/students. It has a total of 72 tokens and 59 types; the total number of distinct words is 1,131 Arabic words. All scripts were written with MSA as the target language and were 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 would appear that all members of the non-native Arabic speaker group are native speakers of English. The corpus includes both male and female speakers.

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, as shown in Fig. 1. 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 the recognition module whose function is to try to figure out the meaning of the input speech given in the testing phase. This module should make the right judgment about what are the best phonemes, syllables, and/or words that were uttered in the input speech. This module can consult system's knowledge inquired from training phase to figure out all possible speech units to select from. This was done with the aid of the HMM models mentioned above.

Table 3. LDC WestPoint Corpus 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 of Speech Files			
	Male	**Female**	**total**
Native	4107	3163	7270
Non-native	883	363	1246
Totals	4990	3526	8516

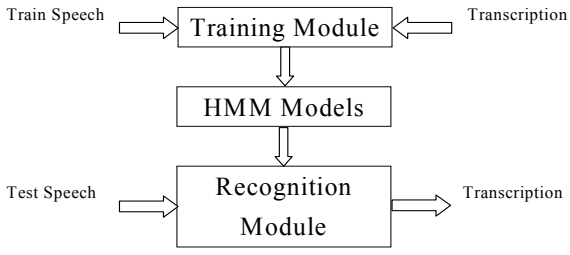

Fig. 1. System block diagram.

As given in Table 4, the parameters of the system were 22 KHz sampling rate with 16 bit sample resolution, 25 millisecond Hamming window duration with a step size of 10 milliseconds, 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.

Table 4. System parameters

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

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 (Linguistic Data Consortium, 2002). The LDC phoneme list as described by LDC WestPoint (Linguistic Data Consortium, 2002) is shown in Table 5 along with the corresponding IPA symbolization. We note that the WestPoint Corpus contains more monophones than the number of MSA phonemes mentioned in the linguistic literature (Omar, 1991; Alkhouli, 1990; Elshafei, 1991). Specifically, WestPoint has added three more phonemes: /g/ "voiced velar plosive", /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. Unfortunately, LDC did not provide any details about the native language of those speakers and other languages that they might master. 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.

Table 5. LDC phoneme list as described by LDC WestPoint (Linguistic Data Consortium, 2002)

LDC Phoneme	Description	IPA Symbol	LDC Phoneme	Description	IPA Symbol
C	voiced pharyngeal fricative	ʕ	ih	high front lax vowel	i
D	velarized voiced alveolar stop	ɖˤ	iy	high front tense vowel	ii
G	voiced velar fricative	ɣ	j	voiced palato-alveolar fricative	ʤ
H	voiceless pharyngeal fricative	ħ	k	voiceless velar stop	k
Q	voiceless glottal stop	ʔ	l	voiced alveolar lateral	l
S	velarized voiceless alveolar fricative	sˤ	m	voiced bilabial nasal	m
T	velarized voiceless alveolar stop	tˤ	n	voiced alveolar nasal	n
TH	velarized voiced interdental fricative	ðˤ	q	voiceless uvular stop	q
Z	voiced interdental fricative	ðˤ	r	voiced alveolar flap	r
ae	low front vowel	aa	s	voiceless alveolar fricative	s
ah	low back vowel	a	sh	voiceless palato-alveolar fricative	ʃ
aw	back upgliding diphthong	aw	t	voiceless alveolar stop	t
ay	front upgliding diphthong	ai	th	voiceless interdental fricative	θ
b	bilabial voiced stop	b	uw	high back rounded vowel	u
d	voiced alveolar stop	d	w	voiced bilabial approximant	w
ey	upper mid front vowel	ay	x	voiceless velar fricative	x
f	voiceless labiodental fricative	f	y	voiced palatal approximant	j
g	voiced velar stop	g	z	voiced alveolar fricative	z
h	voiceless glottal fricative	h			

3. Results and Discussion

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. N/N indicates that native Arabic speakers are used in both training and testing phases, NN/NN implies that non-native Arabic speakers are used both in the training and the test, and M implies that a mixture of native and non-native Arabic speakers is used. As expressed by this terminology, native Arabic speakers were used in both training and testing data of the N/N experiment. On the other hand, native Arabic speakers were used in training data while non-native Arabic speakers were used in testing data of the N/NN experiment. Regarding the NN/N experiment, non-native Arabic speakers were used in training data, while native Arabic speakers were used in testing data. Without using any native Arabic speakers, non-native Arabic speakers were 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 training data and testing data subsets in any given experiment were completely disjoint. In addition to this, the percentage of different sounds, genders, and ages were taken in consideration. We used all male and female speakers and audio files as described in Table 3.

The results are presented in four subsections. The first subsection reports the accuracies for the emphatic sounds and draws some preliminary conclusions. The second subsection presents the same information for the non-emphatic counterparts of the four emphatic sounds. The third subsection analyses all eight target sounds in the time and frequency domains. The last subsection is a general discussion based on the observations.

3.1. Emphatic consonant recognition

Figure 2 plots the accuracies of the four Arabic emphatic consonants for all five experiments. The accuracies for emphatic /D/ are 79.6%, 71.4%, 9.7%, 14.1% and 63.2% in experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. The best accuracy for this phoneme was achieved when using native Arabic speakers in both the training and testing data of the recognition system, i.e. in the N/N experiment. On the other hand, the poorest accuracy was found when non-native Arabic speakers were used in training the system, with either native or non-native Arabic speakers used for testing, i.e. in the NN/N and NN/NN experiments. It is clear from Fig. 2 that the phoneme /D/ achieved relatively poor performance whenever only non-native Arabic speakers were involved in training the system.

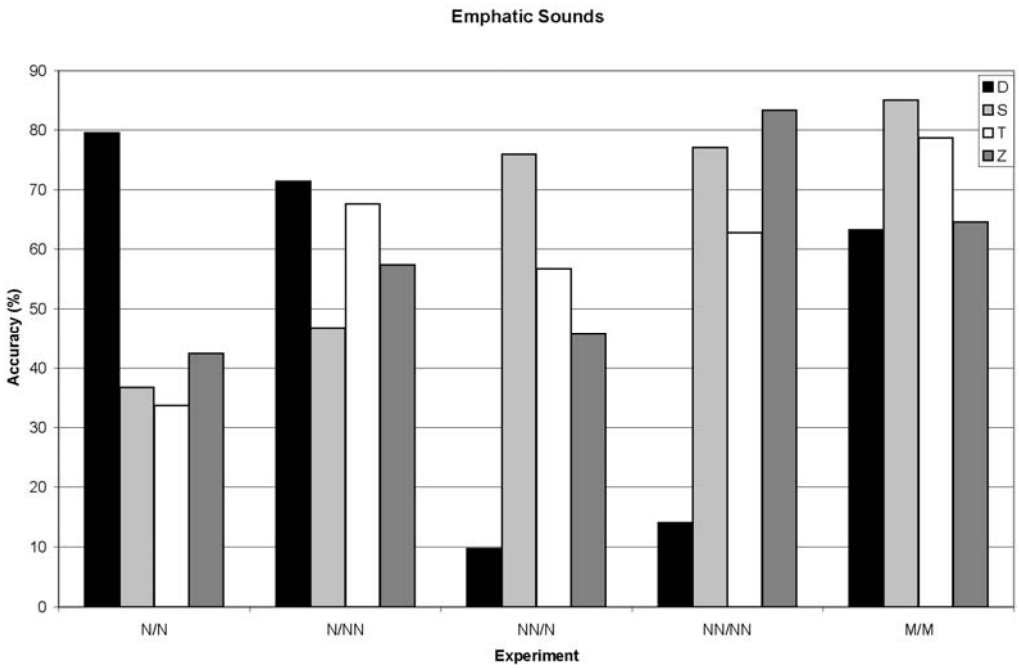

Fig. 2. Performance of emphatics.

Based on this result, we can say that non-native Arabic speakers cannot pronounce /D/ correctly and, hence, that they cannot be used to train the recognition system for this specific phoneme. In other words, non-native speakers are going to give the recognition system misleading knowledge regarding the /D/ phoneme. Indeed, those who have called Arabic "the *Dhaad* language" are absolutely correct because this name implies that non-native Arabic speakers have considerable difficulty pronouncing this phoneme correctly, as we have found here. The term *Dhaad* is the spoken Arabic alphabet letter that carries the /D/ phoneme. When the system was trained by native Arabic speakers or by a mixture of native and non-native Arabic speakers, the accuracy of the /D/ phoneme is relatively high.

The accuracies for emphatic /S/ are 36.7%, 46.8%, 76.0%, 77.1% and 85.1% in experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. The best accuracy for phoneme /S/ was found in experiment M/M, when both training and testing data by the recognition system contained both native and non-native Arabic speakers. In this case, the accuracy was 85.1%. On the other hand, the accuracy for this sound dropped to 36.7% when only native Arabic speakers were used in both training and testing data of the recognition system (i.e., in the N/N experiment).

In contrast to emphatic /D/, the emphatic /S/ sound received a good accuracy whenever the system was trained by non-native Arabic speakers, as shown by the performance in the NN/N and NN/NN experiments. We listened to samples of recorded sentences from the corpus that contain /S/ sounds pronounced by both native and non-native Arabic speakers, and we noticed that, in comparison to native speakers, non-native Arabic speakers gave greater articulatory stress and seemed to pay more attention to the /S/ sound. This may explain the opposite results for this sound as compared to the /D/ sound. It is noteworthy that while stress and careful articulation gave good results in the case of this relatively easy phoneme /S/, no gain from such extra efforts by non-native Arabic speakers could improve results for the /D/ phoneme. By consulting the confusion matrix for the N/N experiment, we found that /S/ was confused most often with its non-emphatic counterpart /s/ and with vowels. This implied that, to the recognition system, /S/ looks like the /s/ sound and like vowels rather than itself.

The /T/ consonant received accuracies of 33.7%, 67.6%, 56.7%, 62.8% and 78.7% in experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. The worst accuracy for this phoneme was noticed in the N/N experiment, where native Arabic speakers were used in both the training and testing data of the recognition system. On the other hand, the best accuracy was found in the M/M experiment where a mixture of native and non-native speakers was used in both training and testing data. Generally speaking, if we exclude the results of the N/N experiment which gave the worst accuracy for recognizing /T/, the /T/ phoneme seems to be less sensitive to speakers' mother tongue. This conclusion is supported by the accuracies of this sound in the other four experiments; all were high with no big differences among them. In the N/N experiment, this sound was mostly confused with the phonemes /Q/, /g/ and vowels.

Accuracies for the empathic sound /Z/ were 42.4%, 57.4%, 45.7%, 83.3% and 64.6% in experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. The worst case for this phoneme was found in the N/N experiment where native Arabic speakers were used in both training and testing of the system. Checking the confusion matrix for the N/N experiment, it was found that this sound was confused mostly with /Q/ and the vowel /ih/. The best accuracy for this sound was encountered with experiment NN/NN (i.e., when non-native Arabic speakers was used in both training and testing data of the system). These results for the sound /Z/ lead us to suggest that non-native speakers can train the system correctly for the empathic /Z/ sound. We note that, as was the case with the /T/ phoneme, the /Z/ phoneme is not sensitive to the mother tongue of the speakers.

We observe that three of the four Arabic emphatic sounds receive better accuracies in those experiments where the recognition system was trained and tested by using a mixture of both native and non-native Arabic speakers (i.e., with experiment M/M). We suggest that in these specific cases the effect of speakers' mother tongue was neutralized.

3.2. Non-emphatic counterparts

In this subsection, we present the accuracies of the non-emphatic counterparts of the four emphatic consonants. We proceed in the same manner as we did in the previous subsection. Figure 3 depicts the accuracies for these sounds in all five experiments.

The Arabic sound /d/, the non-emphatic counterpart of the /D/ phoneme, received the following accuracies: 14.5%, 5.8%, 67.0%, 67.8% and 8.5% for experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. These results are very surprising! They suggest that the system gives very poor results whenever native Arabic speakers were involved in training (i.e., in experiments N/N,

N/NN and M/M). To state this in other words, the /d/ phoneme will be learned more effectively by non-native Arabic speakers, such as shown in experiments NN/N and NN/NN. Our explanation for this phenomenon is as follows: the non-native Arabic speakers pronounce the phoneme /d/ more carefully than native Arabic speakers. Data from the confusion matrix for this sound showed that the /d/ phoneme was mostly confused with its emphatic counterpart, with /Q/, and with vowels.

For the /s/ sound, which is the non-emphatic counterpart of /S/, the following accuracies were observed: 79.7%, 50.2%, 36.4%, 44.1% and 27.6% for experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. The best accuracy for this sound was in experiment N/N where the native Arabic speakers was used in both training and testing data of the recognition system. On the other hand, the poorest accuracy was encountered in experiment M/M, where a mixture of speakers was used in both the training and testing data of the recognition system. Consulting the related confusion matrix for the M/M experiment, it was found that this sound was confused most often with its emphatic counterpart /S/ and with vowels.

Regarding the sound /t/ which is the non-emphatic counter part of /T/, the system accuracies were as follows: 61.2%, 55.4%, 13.5%, 27.2% and 47.0% for experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. The lowest accuracy for this phoneme was shown whenever the recognition system trained with non-native Arabic speakers as in experiments NN/N and NN/NN. Otherwise, the accuracies of the system were significantly better. Regarding the worst case of /t/ accuracy which happened in experiment NN/N (non-native Arabic speakers for training data and native Arabic speakers for testing data of the recognition system), this sound was mostly confused with phonemes /Q/, /T/, /TH/, /d/, /k/, /r/ and vowels.

The sound /TH/, which is the non-emphatic counterpart of /Z/, received the following accuracies: 58.9%, 36.0%, 59.7%, 75.9% and 86.5% for experiments N/N, N/NN, NN/N, NN/NN and M/M, respectively. The poorest accuracy was encountered in experiment N/NN where native Arabic speakers were used in training data of the system and non-native Arabic speakers were used in the testing phase of the system. In this situation, this phoneme was confused mostly with phonemes /f/, /r/, /Q/ and /g/. The best accuracy was found in experiment M/M. Here, a mixture of native and non-native Arabic speakers was used in both training and testing data of the recognition system.

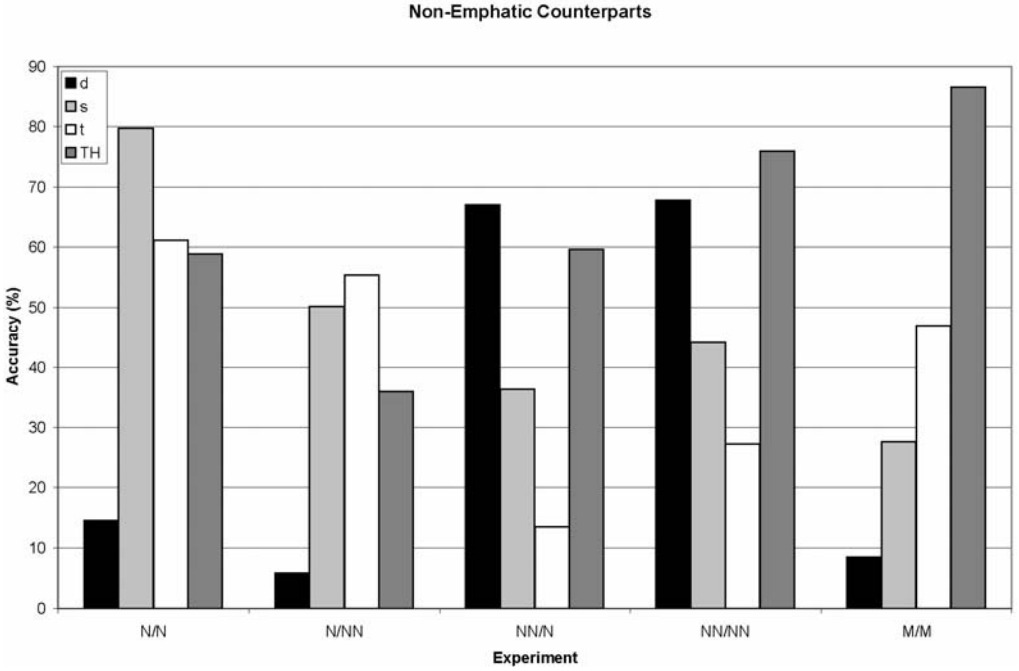

Fig. 3. Performance of the non-emphatic counterparts.

3.3. Emphaticness in the frequency domain

The Spoken Arabic alphabet letters that carry the Arabic emphatic phonemes /D/, /S/, /T/ and /Z/ are *Dhaad*, *Saad*, *T_aa*, and *Dhaa*, respectively. The pronunciations of these spoken carrier words are given (in phonemic representation) as follows: /D a: d/, /S a: d/, /T a:/ and /Z a:/, respectively. In addition, the non-emphatic counter parts of these emphatic Arabic sounds—/d/, /s/, /t/ and /TH/—have the following spoken Arabic alphabet letter carriers: *Daal*, *Seen*, *Taa* and *Thaa*, respectively. These are pronounced as: /d a: l/, /s a: n/, /t a:/ and /TH a: l/, respectively.

The plots of waveforms and spectrograms for the four emphatic consonants and their non-emphatic counterpart are given in Figs. 4 to 7. All speech utterances used in these figures were recorded by native Arabic speakers, not from LDC WestPoint speakers set. The following comparisons are shown: *Dhaad* and *Daal* in Fig. 4, *T aa* and *Taa* in Fig. 5, *Saad* and *San* in Fig. 6, and *Dhaa* and *Thaal* in Fig. 7. The emphatic consonant is given in the upper part of each figure and its non-emphatic counterpart in the lower part.

Brief comparisons of the acoustic features of the emphatic and non-emphatic sounds are as follows:

- In both pairs of plosive consonants, the boundary between emphatic consonant and vowel appears as a sharper spike than in the unemphatic consonant-vowel boundary.
- /D/ shows no delay in the start of the following vowel while /d/ has a delay of about 15 msec; /T/ has a shorter voice onset time than /t/ in the order of 10 msec vs 40 msec.
- The emphatic /S/ fricative shows less intense frication than its non-emphatic counterpart /s/; random noise starts at approximately 4000 Hz in the case of /S/ and at about 3500 Hz in the case of /s/; similarly, /Z/ shows less intense frication than non-emphatic counterpart /TH/.
- The vowel following each of the emphatic consonants has greater concentrations of intensity in the lower frequency range when compared with the vowel following the non-emphatic consonant; in vowels following the emphatic consonant, the F2 is in general lower corresponding to a backing of the low vowel.

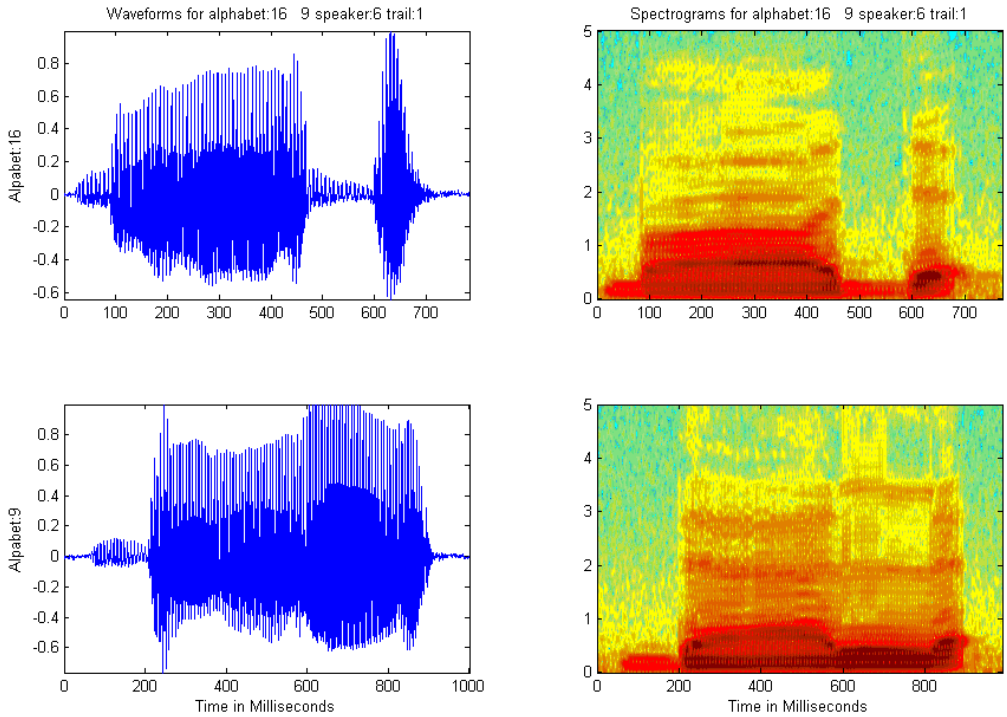

Fig. 4. *Dhaad* and *Daal*.

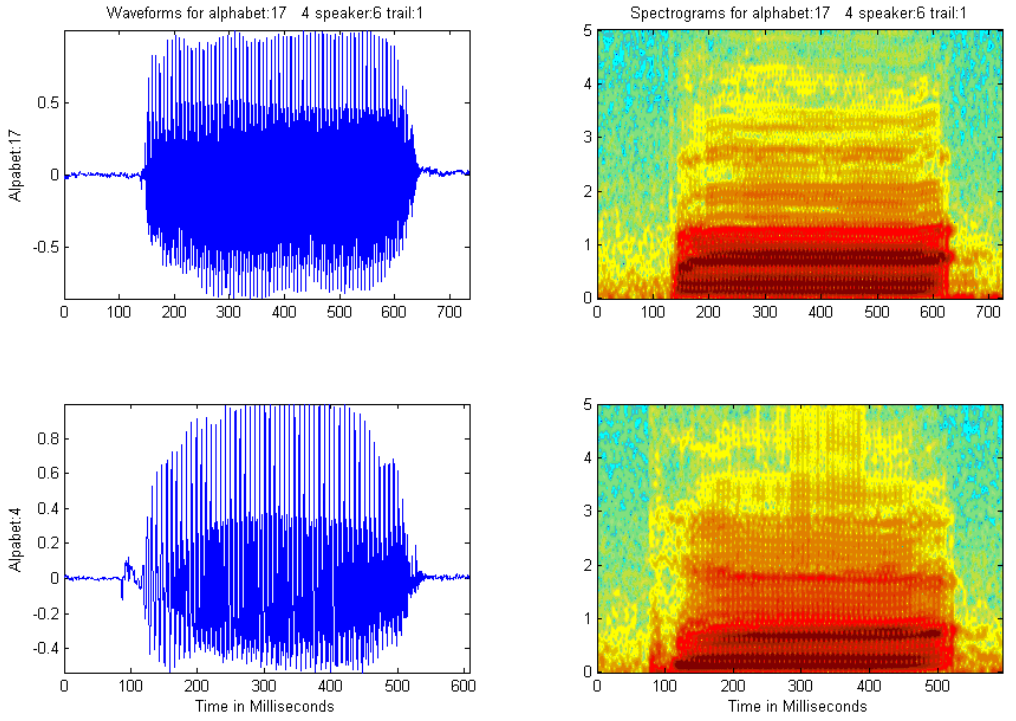

Fig. 5. *T aa* and *Taa*.

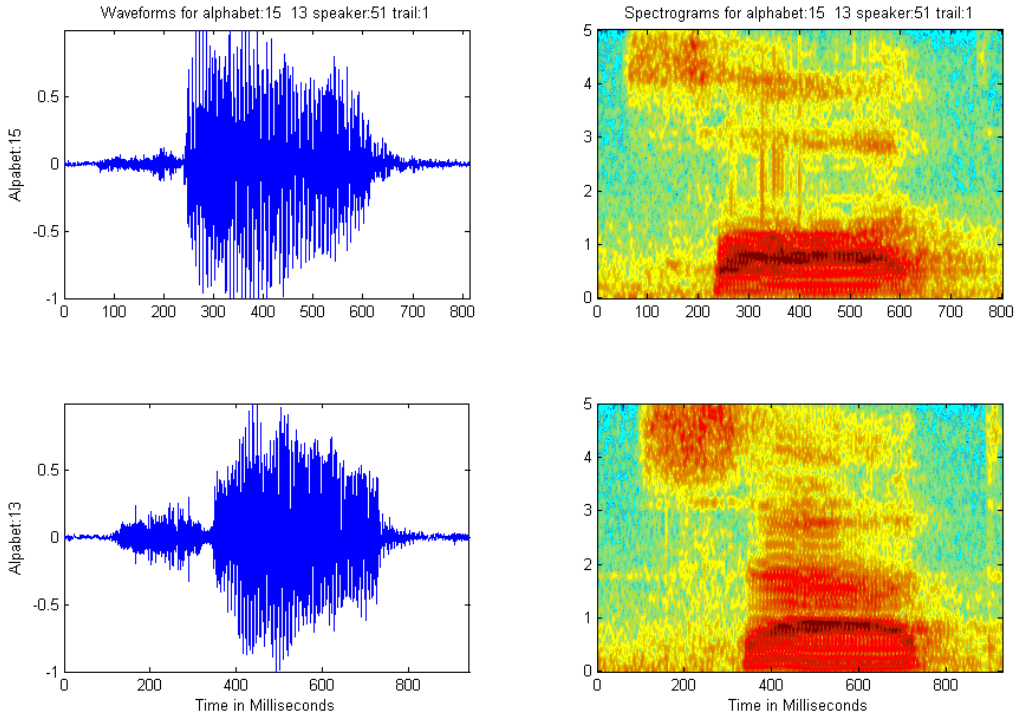

Fig. 6. *Saad* and *San*.

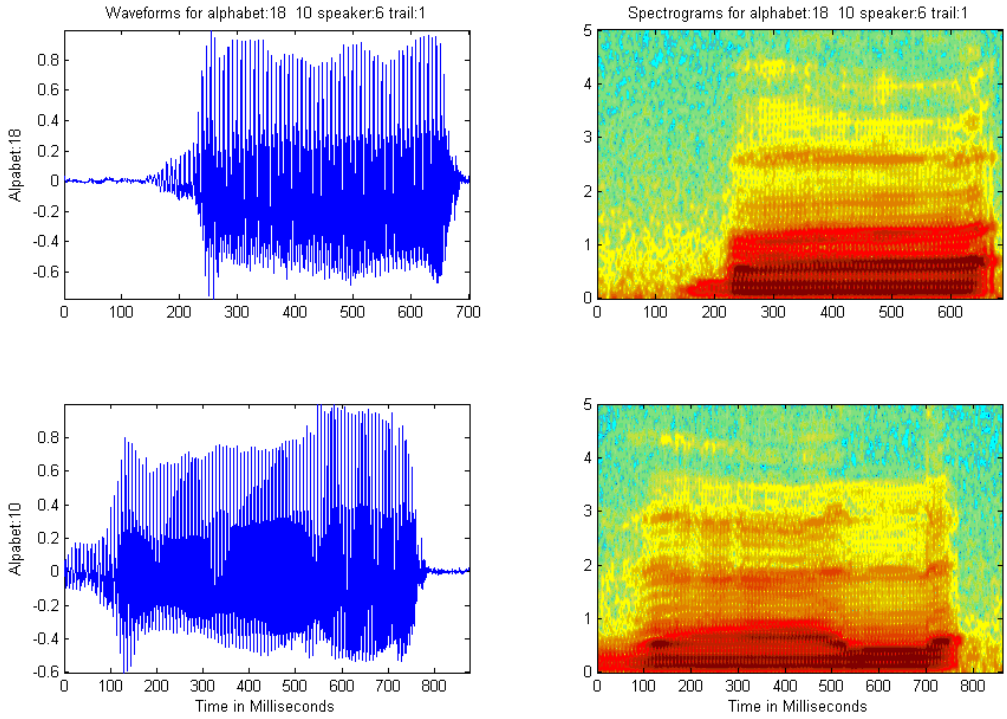

Fig. 7. *Dhaa* and *Thaal*.

3.4. Discussion

The results of the five experiments show the complex difficulties that the four emphatic consonants pose for Arabic ASR. Looking specifically at the N/N experiment, the /D/ phoneme received relatively good accuracies but the system performed less well for the other three consonants—/T/, /S/ and /Z/. The opposite situation obtained for the corresponding non-emphatic consonants where /d/ received poor accuracies compared with /t/, /s/ and /TH/. These results of the N/N experiment reproduce findings of earlier Arabic ASR studies.

The addition of non-native speakers to the research provides several advantages for ASR, as well as some disadvantages. The M/M experiment shows a general improvement in the overall accuracies for /T/, /S/ and /Z/, along with a reasonable accuracy for /D/. Similarly, /TH/ received good accuracy, although /t/, /d/ and /s/ did not. It would appear that certain sounds that exist in English, the native language of the non-native speakers, may provide a source of interference for ASR. This is the case of /t/, /d/ and /s/. However, the careful articulation and attention paid by the non-native speakers to the emphatic consonants, which are not found in English, may well be the source of improvement in the accuracy rates. Nevertheless, the /D/ consonant causes considerable difficulty for non-native speakers of Arabic whose native language is English.

One of the main sources of confusion for the ASR system is associated with vowels. In all five experiments, the system tended to "mis-recognize" the emphatic consonants as vowels. Several explanations can be proposed. First, we note some disagreement with respect to the definition and number of vowels in MSA. We have already mentioned some inconsistencies between the LDC WestPoint Corpus labels and the phonemes given by many Arabic linguists. For example, the long vowel /u:/, which is common in MSA, is not present in the LDC Corpus. The disagreement may have been inspired by the effect of Classical Arabic and of regional Arabic dialects, and perhaps by the vowels in the first language of the non-native speakers. In addition, impressionist aural inspection provided an important clue. Our careful listening to a number of the WestPoint Corpus audio files revealed considerable variation in the pronunciations of vowels by the native Arabic speakers. Our (YAA and S-AA) experience as native speakers of Arabic suggested that this variability in vowel pronunciation

is associated with speakers' regions of origin. Indeed, the MSA in the Corpus has a foreign-accented quality. However, the LDC documentation does not provide detailed information about the regional origins of the speakers. Regional accent can be an important source of confusion for the recognition system, and controlling for phonetic variation in MSA due to region is called for. Sociolinguistic investigations of phonetic variation have established that regional accent can have an effect on vowels, in addition to other factors such as gender and age (Chambers, 1995). For a number of reasons, vowels are part of the difficulty posed for the recognition of the four emphatic consonants.

Another source of difficulty for the ASR system is likely found in co-articulation. That is, the effect that a surrounding phoneme can have on the pronunciation of a target sound. For example, a non-emphatic sound may take on an emphatic quality due to the presence of a neighboring emphatic consonant. Thus, although there are four emphatic phonemes in Arabic, other non-emphatic phonemes may be pharyngealized (the name of the emphatic quality) due to the existence of a neighboring emphatic sound. ASR system errors may be caused by this factor. In studies of several Arabic dialects, the effect of pharyngealization has been found to spread beyond the immediate neighboring segment to other segments in the word. The spread can be in different directions (left and right) and in different domains such as syllable and word (Watson, 1999). Related processes such as labialization are also found in Arabic dialects. The role of co-articulation and spreading in MSA and their effects on Arabic ASR remains to be determined.

4. Conclusion

An Arabic phoneme recognition system was designed and used to investigate four emphatic consonants and their non-emphatic counterparts in Modern Standard Arabic. Five experiments involving both native and non-native speakers of Arabic confirmed the difficulty of these phonemes for ASR. The inclusion of the non-native speakers in the training and testing stages of the system provided several advantages. While these speakers cannot be used to train the recognition system in the case of the emphatic /D/ phoneme because they have considerable difficulty pronouncing this sound, they do provide an advantage for the training of other emphatic phonemes such as /T/, /S/ and /Z/. Frequency domain analyses of both the emphatic and the non-emphatic consonants were discussed.

The paper noted several significant directions for future research. The investigation pointed out discrepancies between phoneme inventories supplied by the LDC WestPoint Corpus and those used by Arabic linguists. 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.

References

Abdulah, W. and Abdul-Karim, M. "Real-time Spoken Arabic Recognizer." *Int. J. Electronics*, Vol. 59, No. (5), (1984), 645-648.

Al-Ani, S.H. *Arabic Phonology: An Acoustical and Physiological Investigation*. Mouton: The Hague, (1970).

Alghamdi, M. *Analysis, Synthesis and Perception of Voicing in Arabic*. Riyadh: Al-Toubah Bookshop, (2004).

Alghamdi, M. *Arabic Phonetics*. Riyadh: Al-Toubah Bookshop, (2001) (in Arabic).

Alkhouli, M. *Alaswaat Allughawiyyah* (Speech Sounds). Jordan: Daar Alfalah, (1990) (in Arabic).

Al-Muhtaseb, H.; Elshafei, M. and Alghamdi, M. "Techniques for High Quality Arabic Text-to-speech." *The Third Workshop on Computer and Information Sciences*, Dammam, Saudi Arabia, (2000), 73-83.

Al-Otaibi, A. *Speech Processing*. The British Library in Association with UMI, (1988).

Al-Zabibi, M. *An Acoustic-phonetic Approach in Automatic Arabic Speech Recognition*. The British Library in Association with UMI, (1990).

Benyettou, A. "The ARABEX Speech Recognition System." *Proceedings of the 1995 IEEE IECON 21st International Electronics, Control, and Instrumentation*, Vol. 2, Orlando, USA, (November 1995), 1068-72.

Chambers, J.K. *Sociolinguistic Theory: Linguistic Variation and Its Social Significance*. Cambridge, MA: Blackwell, (1995).

Cole, R.; Fanty, M.; Muthusamy, Y. and Gopalakrishnan, M. "Speaker-independent Recognition of Spoken English Letters." *International Joint Conference on Neural Networks (IJCNN)*, Vol. 2, San Diego, USA, (June 1990), 45-51.

Cosi, P.; Hosom, J. and Valente, A. "High Performance Telephone Bandwidth Speaker Independent Continuous Digit Recognition." *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Trento, Italy, (2001).

Davis, S. and Mermelstein, P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences." *IEEE Trans. on Acoustic, Speech, and Signal Processing*, Vol. ASSP-28, No. (4), (August 1980).

Deller, J.; Proakis, J. and Hansen, J.H. *Discrete-time Processing of Speech Signal*. Macmillan, (1993).

Economic and Social Commission for Western Asia, United Nations Report. *Harmonization of ICT Standards Related to Arabic Language Use in Information Society Applications*. New York: United Nations, (2003).

El-Imam, Y.A. "An Unrestricted Vocabulary Arabic Speech Synthesis System." *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. 37, No. (12), (December 1989), 1829-45.

El-Imam, Y.A. "Synthesis of Arabic from Short Sound Clusters." *Computer Speech & Language*, Vol. 15, No. (4), (2001), 355-380.

Elobied, A.; Iman, A. and Soghayroun, A. "On Obtaining Parameters for a Model of Arabic Speech Production." *IEEE International Conference on Telecommunication 1994, Bridging East & West Through Communications*, Dubai, UAE, (1994).

Elshafei, M. "Toward an Arabic Text-to-speech System." *The Arabian Journal for Science and Engineering*, Vol. 16, No. (4B), (October 1991), 565-83.

Hagos, E. "Implementation of an Isolated Word Recognition System." *UMI Dissertation Service*, Dhahran, Saudi Arabia, (1985).

Haykin, S. *Neural Networks: A Comprehensive Foundation*. 2nd ed., Prentice Hall, (1999).

Juang, B. and Rabiner, L. "Hidden Markov Models for Speech Recognition." *Technometrics*, Vol. 33, No. (3), (August 1991), 251-272.

Karnjanadecha, M. and Zahorian, Z. "Signal Modeling for High-performance Robust Isolated Word Recognition." *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. (6), (September 2001), 647-654.

Kirchhoff, K.; Bilmes, J.; Das, S.; Duta, N.; Egan, M.; Gang, J.; Feng, H.; Henderson, J.; Daben, L.; Noamany, M.; Schone, P.; Schwartz, R. and Vergyri, D. "Novel Speech Recognition Models for Arabic: Johns-Hopkins University Summer Research Workshop 2002, Final Report." http://www.clsp.jhu.edu/ws02/, (2002).

Kirchhoff, K.; Bilmes, J.; Das, S.; Duta, N.; Egan, M.; Gang, J.; Feng, H.; Henderson, J.; Daben, L.; Noamany, M.; Schone, P.; Schwartz, R. and Vergyri, D. "Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Summer Workshop." *Proceedings of ICASSP 2003*, Vol. 1, Hong Kong, (April 2003), 344-347.

Laufer, A. and Baer, T. "The Emphatic and Pharyngeal Sounds in Hebrew and Arabic." *Language and Speech*, Vol. 31, No. (2), (1988), 181-205.

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

Lippmann, R. "Review of Neural Networks for Speech Recognition." *Neural Computation*, MIT Press, (1989), 1-38.

Loizou, P.C. and Spanias, A.S. "High-performance Alphabet Recognition." *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. (6), (November 1996), 430-445.

Nocerino, N.; Soong, F.; Rabiner, L. and Klatt, D. "Comparative Study of Several Distortion Measures for Speech Recognition." *Speech Communication*, Vol. 4, (1985), 317-31.

Omar, A. *Derasat Alswaut Aloghawi*. Egypt: Aalam Alkutob, (1991) (in Arabic).

Ouni, S.; Cohen, M. and Massaro, W. "Training Baldi to be Multilingual: A Case Study for an Arabic Badr." *Speech Communication*, Vol. 45, (2005), 115-37.

Rabiner, L.R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE*, Vol. 77, No. (2), San Francisco, USA, (February 1989), 257-286.

Rabiner, L. and Juang, B. *Fundamentals of Speech Recognition*. Prentice Hall, (1993).

Rabiner, L. and Samber, M. "An Algorithm for Determining the Endpoints of Isolated Utterances." *The Bell System Technical Journal*, Vol. 54, No. (2), (1975), 297-315.

Rabiner, L. and Wilpon, J. "A Simplified, Robust Training Procedure for Speaker Trained Isolated Word Recognition Systems." *J. Acoustic Society of America*, Vol. 68, No. (5), (November 1980).

Selouani, S. and Caelen, J. "Arabic Phonetic Features Recognition Using Modular Connectionist Architectures." *Interactive Voice Technology for Communication (IVTTA'98), Proceedings 1998 IEEE 4th Workshop*, Torino, Italy, (29-30 September 1998), 155-160.

Watson, J.C.E. "Emphasis in San'ani Arabic." In: *Three Topics in Arabic Phonology*. University of Durham Center for Middle Eastern and Islamic Studies Occasional Papers, Vol. 53, (1996), 45-52.

Watson, J.C.E. "The Directionality of Emphasis Spread in Arabic." *Linguistic Inquiry*, Vol. 30, (1999), 289-300.

Welch, J. "Combination of Linear and Nonlinear for Isolated Word Recognition (Abstract)." *J. Acoustic Society of America*, Suppl. 1, Vol. 67, (Spring 1980), S14.

Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V. and Woodland, P. *The HTK Book (for HTK version 3.3)*. Cambridge University Engineering Department, (2005), http://htk.eng.cam.ac.uk/prot-doc/ktkbook.pdf.

دراسة الصوامت المفخمة في الكلام الأجنبي للعربية

يوسف بن عجمي العتيبي[1]، وسيد أحمد سلواي[2]، ولاديزلا سيشوكي[3]

[1] *قسم هندسة الحاسب، كلية علوم الحاسب والمعلومات، جامعة الملك سعود، الرياض، المملكة العربية السعودية*

[2] *معمل LARIHS، جامعة مونكتون، حرم شيباغان الجامعي، كندا*

[3] *قسم اللغة الفرنسية، جامعة نيو بيرنزويك، كندا*

(قدم للنشر في ٢٠٠٧/١١/٢١م؛ وقبل للنشر في ٢٠٠٩/١/٢٥م)

ملخص البحث. تم في هذا البحث دراسة أربعة صوامت مفخمة عربية من زاوية الأنظمة الآلية للتعرف على الكلام. تمت مقارنة معدلات الخطأ لهذه الصوامت وكذلك لنظائر هذه الصوامت غير المفخمة في خمسة تجارب، وتم تحليل النتائج. تختلف هذه التجارب فقط في عينات التدريب والاختبار من حيث اللغة الأم للمتكلم أهي العربية أم لا، كذلك تم وصف هذه الأصوات في نطاقي الزمن والتردد. استخدمت جميع التجارب برنامج نموذج ماركوف الخفي المعروف بـ(HTK)، والذخيرة الصوتية المسماة وستبوينت من (LDC) والتي كونت خصيصاً للغة العربية الفصحى المعاصرة (MSA). برهنت النتائج على أن الأصوات العربية المفخمة هي مصدر لعجز النظام الآلي للتعرف على الكلام العربي. مع أن صوت الضاد العربي قد حقق معدل نجاح متدنٍّ أقل من ١٥٪، إلا أنه توجد مزية في أداء النظام عند تضمين الكلام العربي للمتحدث غير العربي. إن التغير في إصدار الأصوات العربية الذي يعتمد على اختلاف المنطقة لهو جدير بالاهتمام في أنظمة معالجة الكلام العربي والتعرف عليه آلياً.