Pronunciation Evaluation Model for None Native English Speakers

[1]Hassanin M. Al-Barhamtoshy, [2]Sherif M. Abdou and [1]Kamal M. Jambi

[1]Faculty of Computing, King Abdulaziz University (KAU), Jeddah, Saudi Arabia
[2]Faculty of Computers and Information, Cairo University, Cairo, Egypt
hassanin@kau.edu.sa, s.abdou@fci-cu.edu.eg, kjambi@kau.edu.sa

Abstract: This paper investigates both phonetic and phonological influences of speaker's none native language on their accent in English. Phonological influences of vowels and consonants over the speech data for the *Speak Correct* system will be studied. The Speak Correct layout will be presented in short description. Two groups were involved, whose native language was Arabic, and dialect spoken in Saudi Arabia and Egypt. In Speak Correct system evaluation, participants were asked to utter 17 English vowels and 24 consonants. Both Saudi group and Egyptian group demonstrated accuracy in identification of vowels (70-85 %) and consonants (80-90 %). A testing dataset are described using pronunciation scoring method and experimental assessment for evaluation. Therefore, the paper introduces to test the *Speak Correct* system to pronounced English word.
[Hassanin M. Al-Barhamtoshy, Sherif M. Abdou and Kamal M. Jambi. **Pronunciation Evaluation Model for None Native English Speakers.** *Life Sci J* 2014;11(9):216-226]. (ISSN:1097-8135). http://www.lifesciencesite.com. 30

Keywords. Speech recognition, English vowels and consonants, Arabic dialects, acoustic error.

1. Introduction

English pronunciation affected by variety of factors, ability to accurately perceive the speech sounds of language [1], and different dialects of the same language [2]. English dialects are frequently used in different domains; business, education, governmental, etc [3]. The presence of English need to understand it, and there are increasing to interact with governmental and business sectors using English [1] [4]. Three main factors are playing in English perception [1]: listeners' native language, 920 English dialects, and phonetic context of speech sounds are presented.

1.1 Key Definitions

Phoneme: the smallest unit of speech is a phoneme; which is used to distinguish meaning. The phoneme is the most important unit in a word, each word consists of phonemes, and substituting phonemes causes a change in the meaning of a word. For instance, if the sound [b] is replaced by [p] in the word "pin", the word changed to "bin". Therefore /b/ is a phoneme [5].

Phone: the smallest physical segment of sound. Therefore, phones are the physical realization of phonemes. An **allophone** is a phonic variety of a phoneme [6, 7].

Phonetics: the study of human speech is concerned with the properties of speech sounds.

Phonology: is used to study sound systems and abstract sound units; i.e., phonemes and phonological rules. Therefore, phonetics definitions apply across languages, and phonology is language based. The phonetic meaning of a sound is described using phonology [5, 7], and a phoneme is represented using //.

Syllable: is defined as a unit of pronunciation. It is generally larger than a single sound and smaller than a word. Syllables generally start and end with consonants, and contain vowels.

1.2 Pronunciation

This section illustrates how pronunciation can vary, and how phonemes can have various allophones in different phonetic environments. This section also describes a technique for writing transducer rules to model such changes in speech, including inaccurate accents, specific pronunciation errors, and common pronunciation errors.

Lexical variation and allophonic variation are two classes of pronunciation variation. Lexical variation is used to represent a word in a spoken lexicon, while allophonic variation refers to differences in how individual segments change value [5]. Most pronunciation variation is allophonic, according to the influence of surrounding sounds and syllable structure. Also, the lexical variation is related to sociolinguistic variation, which is caused by extra linguistic factors, such as accent and dialect variations. Other sociolinguistic differences are due to differences in register or style, rather than dialect. A thoroughly researched example of style-variation is the suffix "-ing" (as in "something"), which can be pronounced "somethin" (without the "g") [5, 7].

The proposed rules of pronunciation are dependent on a complicated set of factors that must be interpreted probabilistically. Most allophonic rules in

[1] http://mepi.state.gov

English can be grouped into types: assimilation, dissimilation, deletion, flapping, vowel reduction, and epenthesis (insertion an extra sound into a word) [5].

Assimilation is a change made to a sound segment to make it more like a neighboring segment, e.g.: dentalization and palatalization. As an example of the palatalization rule is as follows:

$$\begin{Bmatrix} [s] \\ [z] \\ [t] \\ [d] \end{Bmatrix} \Longrightarrow \begin{Bmatrix} [ʃ] \\ [ʒ] \\ [ʧ] \\ [ʤ] \end{Bmatrix} / - \{y\}$$

Deletion is the removal of a sound from a word. The following rule shows how /t/ and /d/ are deleted when they occur before consonants:

$$\begin{Bmatrix} t \\ d \end{Bmatrix} \Longrightarrow \theta / V - C$$

Flapping is a type of sound that occurs when a speaker is speaking quickly, and is more likely to happen at the end of a word. Flapping often impacts vowel reduction. Other studies have discussed spelling error patterns that occur in typed text and speech-recognition [5, 8, and 9]. These include single-error misspellings induced by one the following errors: insertion, deletion, substitution, and transposition.

- **Vowels.** Some English phonemes are equivalent or nearly equivalent to Arabic phonemes, and therefore can be articulated without great difficulties. Some English phonemes may be problematic, the following cause the most confusion [7]:

a. /e/ and /ɪ/ are often confused; for example bit for bet.

b. The two phonemes /ɒ/ and /ɔ:/ are often confused; e.g., cot for caught.

c. The diphthongs /əʊ/ and /eɪ/ are pronounced short, and may be confused with /e/ and /ɒ/; e.g., red for raid.

- **Consonants.** Some English phonemes are equivalent or nearly equivalent to Arabic phonemes, and therefore can be articulated without difficulties. Though some confusion may still arise, few phonemes cause problems. The following comments illustrate examples of such problems:

a. The Arabic letter /g/ is pronounced /g/ in an Egyptian accent, /ʤ/ in a Saudi accent, and sometimes even /j/, according to local dialects.

b. The two letters /v/ and /f/ are often confused, especially in a Saudi accent; e.g., it is a fery nice fillage.

c. The two allophones /p/ and /b/ tend to be used somewhat randomly: I baid ten bense for a bicture.

d. Depending on dialect, /θ/ and /ð/ are pronounced as /t/ and /d/, respectively- especially in an Egyptian accent- I tink dat dey …

e. The rolling of /r/ is voiced with a flap, and Arabic speakers may over pronounce the post-vocalic r; as in car park.

f. Sometimes /g/ and /k/ are confused; especially for dialects that do not include the phoneme /g/, as in goat/coat and bag/bak.

1.3 English Vowels and Consonants

Certain English phonemes cannot pronounce correctly for Arabic accent speakers. They often substitute normal phones with that are closest to them. At testing phase of *Speak Correct* system, a pronunciation phase is carried out to find out the English phonemes that they confused with. Table (1) illustrates 17 vowels with grapheme script, word examples, IPA symbol, and vowels description. Arabic (standard form) is similar to the vowel system in the English, as shown in table (1) and (2). Table 2 shows phonemes script, example of consonants, IPA, and consonant description for English.

1.4 Paper Objectives

The main objective of the present paper was to illustrate an initial evaluation of the *Speak Correct* system, from linguistic point of view. This evaluation depends on linguistically-diverse to identify different proposed English vowels and consonants. Therefore, this study took two participants groups; (1) Saudi dialects group and (2) Egyptian dialects group. So, the objectives of this paper can be abstracted in:

1. Evaluate the overall recognition accuracy of the *Speak Correct* system.
2. Assess error patterns for specific English vowels and consonant.
3. Classify/Examine the acoustic/linguistic level at young-adult students at KAU.

1.5 Related Works

The use of frequent pronunciation error in second language made by L2 learner of Dutch often concern vowel substitution [5]. Therefore, ASR-based confidence measure with phonetic feature is used in such pronunciation errors. Additional several studies have been mentioned in such study [10].

Pronunciation assessment of vowels is investigated in (Joshi et al, 2013) [11] using specific combinations of acoustic American English and Hindi models. Accordingly, suitable trained adapted modules were used to achieve pronunciation scoring systems that predict error patterns uttered by different L1 speakers.

However, the effect of POS on Mandarin speech recognition system has been presented in (Gong et al, 2012) [12]. The word in POS establishes to reduce lexical ambiguity in the proposed language model, in addition to provide some information about pronunciation of heteronyms. So, studying of POS on speech recognition system at lexical and acoustic

levels was presented into language model and pronunciation dictionary.

Building a lexicon for speech recognition by using acoustic data-driven and pronunciation learning methodology is addressed in Lu [13]. The pronunciation lexicon uses transcribed acoustic data and WFST-based EM algorithm.

On the other hand, a proposed framework for unsupervised discovery of pronunciation error patterns has been presented in Wang [14].

The results in [15] show that applying letter to sound system for Romanian conversion. Expert system, decision trees, neural networks, SVM, and pronunciation by analogy are five systems introduced in the paper. The tested data showed that decisions trees and neural networks generate best results.

Algerian dialects are variants of Modern Standard Arabic (MSA) stemming, pronunciation and grammar, [16]. The paper investigates the effect of gender of speakers and regional accents on MSA ASR performance.

An analysis on mispronunciation of computer-aided pronunciation training (CAPT) system studied in Jai [17]. Computational method of obtaining the auditory perceptual distance between phonemes are discussed and investigated.

Also, in Computer-Aided Pronunciation Training (CAPT) is very important to be used in error pattern detection, (Wang et al) [18]. Accordingly, linguistic and pedagogical experience is used by experience teachers of English error pattern detection. Therefore, modeling approach is presented with empirical analysis for CAPT.

Table (1): Common Vowels for Arabic Dialect Evaluation List (Saudi and Egypt).

Script	Word Examples		IPA	Description
IY	sea	happy	/iː/	front close monothong long vowel produced with forward shift of the tongue from the rest position
IH	kit	inside	/ɪ/	front close monothong short vowel produced with forward shift of the tongue from the rest position
EH	dress	square	/ɛ/	front middle monothong vowel produced with forward shift of the tongue from the rest position
EY	face	gate	/eɪ/	front middle diphthong vowel produced with forward shift of the tongue from the rest position
AE	trap	cat	/æ/	front open monothong vowel produced with forward shift of the tongue from the rest position
AA	father	start	/ɑː/	back open monothong long vowel produced with forward shift of the tongue from the rest position
AH	cut	Up	ʌ	Central open monothong short vowel produced with the tongue in neutral or rest position
AX	common	upper	ə	Central middle monothong short vowel produced with the tongue in neutral or rest position. It comes under the name schwa which is the most neutral vowel
ER	nurse	bird	ɜːr	Central middle diphthong long vowel produced with the tongue in neutral or rest position then end with the r sound.
UW	June	room	uː	Back closed monothong long vowel produced with the backward shift of the tongue from its neutral or rest position
UH	foot	put	ʊ	Back closed monothong long vowel produced with the backward shift of the tongue from its neutral or rest position
AO	thought	law	ɔː	Back middle monothong long vowel produced with the backward shift of the tongue from its neutral or rest position
OW	goat	low	oʊ	Back middle open diphthong short vowel produced with the backward shift of the tongue from its neutral or rest position
OH	stop	accommodate	ɒ	Back open short vowel produced with the backward shift of the tongue from its neutral or rest position
AY	price	mine	aɪ	Diphthong of two vowels produced consecutively by moving the articulator from the position of vowel a to vowel I
OY	choice	boy	ɔɪ	Diphthong of two vowels produced consecutively by moving the articulator from the position of vowel ɔ to vowel I
AW	mouth	hour	aʊ	Diphthong of two vowels produced consecutively by moving the articulator from the position of vowel a to vowel ʊ

Table (2): Common Consonants Evaluation List for Arabic Accent (Saudi and Egypt).

Script	Word Examples		IPA	Description
B	back	book	b	Voiced bilabial stop
CH	chair	choose	tʃ	Voiceless post-alveolar affricate
D	day	dear	d	Voiced alveolar stop
DH	this	then	ð	Voiceless dental fricative
F	fat	fear	f	Voiceless labiodental fricative
G	get	go	g	Voiced velar stop
HH	hot	high	h	Voiceless glottal fricative
JH	judge	jury	dʒ	Voiced post-alveolar affricate
K	key	keen	k	Voiceless velar stop
L	light	metal	l	Lateral alveolar
M	more	make	m	Voiced bilabial nasal
N	nice	cotton	n	Voiced alveolar nasal
NG	ring	thing	ŋ	Nasal velar
P	pen	shop	p	Voiceless bilabial stop
R	right	fear	r	Approximate alveolar
S	soon	loose	s	Voiceless alveolar fricative
SH	sure	future	ʃ	Voiceless post-alveolar fricative
T	tea	meet	t	Voiceless alveolar stop
TH	thing	both	θ	Voiced dental fricative
V	vet	save	v	Voiced labiodental fricative
W	wet	where	w	Approximate labio-velar
Y	yet	your	j	Approximate platal
Z	zero	freeze	z	Voiced alveolar fricative
ZH	pleasure	measure	ʒ	Voiced post-alveolar fricative

2. System Architecture

Researchers have introduced many algorithms for use in speech recognition. For instance, algorithms for phone and syllable have previously been described [8, 19, and 20]. In addition, the N-gram language model and the Hidden Markov Model (HMM) have been discussed previously [19, 20].

The HMM was first described as a stochastic method for modeling temporal pattern recognition and sequencing data. Therefore, the HMM can be illustrated using finite state machines: at each transition there is an observation from a specific state, for each state there is an output symbol emission [8]. In other words, to choose a word that is the most probable given an observation, a single word such that P (word | observation) is most likely. If w is the estimated correct word and O is the observed sequence (individual observation), then the equation for picking the best word is given as:

$$W = \text{argmax } P(o|w) \quad P(w)$$

where: P(o|w) represents likelihood, P(*w*) represents prior, w is vocabulary, *w* is the correct word, and *o* is observation. Once the likelihood-computation has been solved, and decoding for a simplified input consisting of strings of phones have been established, feature extraction will quickly be resolved.

2.1. The Speak Correct Model Architecture

The Speak Correct system architecture is illustrated in figure 1. It consists from three main modules. The first module includes the training module, a regression training language model is trained from collection of acoustic model and generic miss-pronunciation model. So, the predefined language model can be delivered through clean speech in order to fed and generate spectra features to the second module (decoder module). The training module uses HMM layers, and starts with the initial acoustic language model as in (Abdou et al) [7]. Such acoustic model is based on feature space Maximum Likelihood Linear Regression (MLLR) using word by word from pre-recorded speech to build standard adaption model.

The middle part of figure 1 contains the decoding module as the seconded part in Speak Correct system. This module is based on the Automatic Speech Recognition (ASR) technology [7]. The algorithm of such technology is related to how to calculate matching score between the speakers' utterances and the acoustic language models. If that score greater than threshold the

speaker utterance will be judged as a correct one, else it will be rejected.

Consequently, speaker utterances are processed by the decoder model to generate phone lattice as data structure's directed graph. This data structure contains hypothesis' grammars related to the phones uttered. Afterwards, it uses an acoustic HMM and phontactic N-gram, which are both trained from training module. The phone lattice is then received with a similar structure that includes a phone-level pronunciation model to capture unpredicted mistakes.

The evaluation module (third module) receives the pronounced sounds through operations on finite state machines, and signals miss pronunciation message if some mistakes is happened. Consequently, such module combines user's utterances with related pronunciation errors, and combined them with the speaker profile to produce users' feedback.

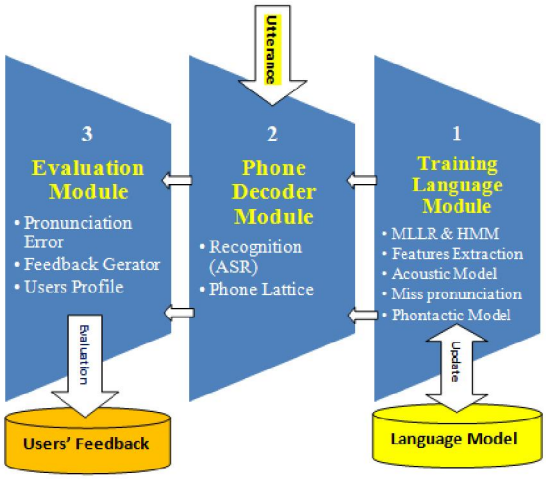

Figure 1: The Architecture of Speak Correct System

2.2. Acoustic Errors Analysis

Therefore, acoustic analysis is used to find out the difference properties between native speakers and non native speakers (Saudi and Egyptian accents). Consequently, comparative analysis will be used to find the difference between phonemes that Arabian speakers potentially be confused with vowels and consonants. One ambiguity of such pronunciation is illustrated in figure (2) that shows the two waveform of the two English words "back" and "pack" that many Arabic speakers especially Arabic speakers be confused with. From such figures voiced features are displayed, there is difference between the beginnings of the two words that include the /p/ on the first phoneme of the word "pack" that indicates substitute /b/ instead of /p/. Also, figure 2 illustrates the difference between the two words at the two levels; waveform and spectral forms- this difference is due to substitution of phone instead of another.

Speech corpus of the *Speak Correct* is prepared using 100 recorded hours by native speakers. The training set contains 70 of Saudi and Egyptian speakers, while the testing set during development includes 30 speakers from both regions. On the other hand, the test phase includes 10 Egyptian and 5 Saudi speakers. As mentioned before, the experiment is carried out using *Speak Correct* system [7].

The language model of the *Speak Correct* is built using such 100 recorded hours from American's Speakers of local news. Table (3) shows the classification of numbers and nationality of speakers during testing phase, taken into consideration to cover all the vowels and consonants phonemes.

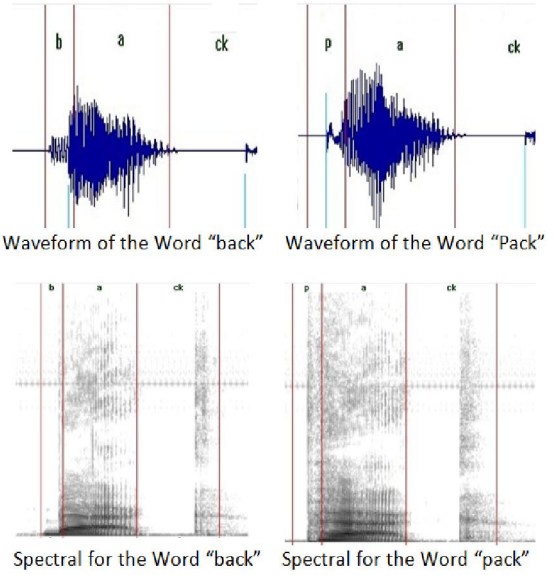

Figure 2: Waveform and Spectral View of Pronounced Words

Table (3): No of Speakers, Nationality with respect to Vowels and Consonants

Nationality	Number	Vowels	Consonants
Saudi	5	17	24
Egyptian	10	17	24

3. Speak Correct Testing

To test *SpeakCorrect* system, the dataset of such testing covers the 17 vowel phones, and 24 consonants; as mentioned before. Most of the words of this dataset contain more than one vowel. As we mentioned in previous literature [21], substitution, deletion and insertion are the most important errors in mispronunciation [20]. There are two levels to cover the Arabic accent mispronunciation; level 1:

Vowel pronunciation and level 2: Consonant pronunciation. At each level the word lists were read out by 10 speakers (students in our case). Most of those students had been studied at faculty of computing and information technology (FCIT) in King Abdulaziz University (KAU). The selected English words at each lesson contained between 97-23 words for each vowel or consonant – shown in tables 1 and 2.

Each student (speaker) read aloud the word in the dataset after selecting level, lesson number and then press record button to read the displayed word. The speech (sound) was recorded using quality microphone with 16 KHz frequency and 32 bit mono wave. Table (4) details such dataset classification.

3.1 Procedurally Method

The vowel test of the proposed *Speak Correct* system was based on the material developed in our project of (INF-1406-03-10), available at faculty of computing, KAU. It includes two levels for evaluation; level 1 contains 17 lessons to handle all the English vowels to analysis all dialects defects and level 2 includes 5 lessons of consonants.

Participants' response was analyzed separately, within each group (Saudi and Egypt). Therefore, each group was obtained accuracy across vowels and consonants in *Speak Correct* testing. Next, recognition accuracy was computed for each individual vowel (17 lessons; and each lessons contains 7-33 vowels' examples), as well as for each individual consonants (5 lessons, and each lesson includes 10-33 consonants' examples). Lastly, confusion errors matrix were measured for each student and combined into confusion tables (4 and 5).

Confusion matrix for the dataset is manually annotated at two levels phone level and consonant level, to obtain surface perceived transcription, as shown in Table 4. This table identifies the most common confusions or mispronunciations speakers in English in Saudi and Egypt regions.

3.2 Vowels

Table 4 contains the actual and predictable classifications by *Speak Correct* system. This table shows the confusion values for the 17 vowels of the *Speak Correct* system. The diagonal elements represent the percentage number of correctly classified vowels during recognition, for which the predictable actual vowel is equal to the perceived vowel. The higher values at the diagonal indicate many correct perceived or predictable values.

The confusion matrix indicates that the common errors include of [ar] vowel for [ax] vowel (42.83 % of error phones), and [ah] vowel for [ow] vowel (14.66 % error phones). Notice that the periods (-) signify cells whose value is 0. Figure 3 illustrates percentage of correct classified vowels relative to actual vowels through recognition test of the *Speak Correct* system.

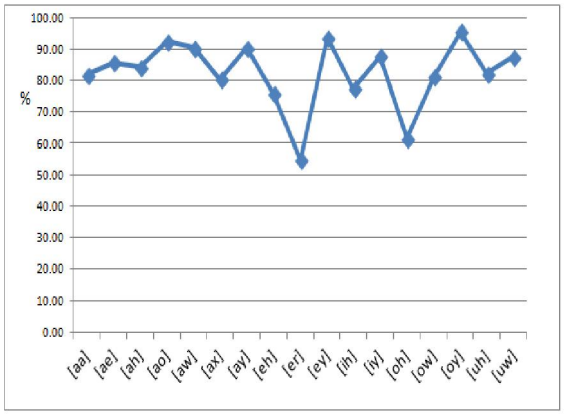

Figure 3: Relation between Accuracy Percentages of Vowels

3.3 Consonants

Overall consonants identification accuracy for the *Speak Correct* system was reasonable, across vocalic words (84.3 %). Also, during testing, the highest number of errors is located for /CH/, which must often miss-identified as /SH/ (45% confused error and 52.70% correct), see Table 5. Additional minor confusions were found between the consonant /V/ which was sometimes identified as /F/ (31.17 %). This distribution of errors may suggest that the /F:V/ confusions were due to the preliminary dialects, especially for Saudi accents, which may be influenced by the phonologies of local regional in Arabic area.

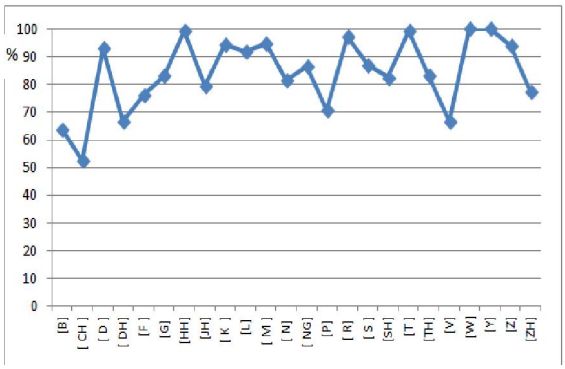

Figure 4: Relation between Accuracy Percentages of Consonants

We use phone error rate (PER) to measure error correction performance which is calculated using equation 1.

$$PER = (|P_s| + |P_i| + |P_d|) / T_w \quad \ldots\ldots\ldots\ldots\ldots (1)$$

T_w represents total number of words in the dataset. $|P_s|$, $|P_i|$ and $|P_d|$ are number of substitution, insertion and deletion errors respectively.

Table (4): Vowels Confusion Matrix

Actual Vowels \ Perceived Vowels	[aa]	[ae]	[ah]	[ao]	[aw]	[ax]	[ay]	[eh]	[er]	[ey]	[ih]	[iy]	[oh]	[ow]	[oy]	[uh]	[uw]
[aa]	81.65	-	1.43	14.99	-	0.16	0.04	-	-	-	0.02	-	-	1.41	0.02	0.11	0.16
[ae]	12.42	85.75	0.26	0.50	0.03	0.71	-	-	-	0.21	0.05	0.05	-	-	-	0.03	-
[ah]	1.24	0.31	84.20	4.28	0.14	5.97	-	-	0.03	-	0.07	0.03	-	0.14	0.03	1.00	2.55
[ao]	0.40	0.23	0.30	92.52	0.26	0.28	-	-	-	0.02	-	-	-	2.87	0.07	2.17	0.89
[aw]	3.93	-	1.81	0.30	90.48	0.60	-	-	-	-	-	-	-	2.57	0.15	-	0.15
[ax]	1.19	1.30	0.11	2.27	0.14	80.28	-	-	-	0.22	9.52	0.11	-	0.47	-	4.04	0.36
[ay]	1.08	0.04	-	0.04	0.22	0.09	90.35	-	-	1.38	4.74	1.98	-	-	-	-	0.09
[eh]	-	22.22	-	-	-	0.44	0.11	75.70	0.04	0.22	0.55	0.55	-	-	-	-	0.18
[er]	0.16	0.02	0.25	0.47	0.24	**42.83**	0.27	-	**54.69**	0.18	0.25	0.13	-	0.16	0.02	0.13	0.20
[ey]	0.17	0.37	0.09	0.06	-	0.32	0.40	-	0.03	93.73	4.76	0.03	-	0.03	0.03	-	-
[ih]	-	-	0.02	0.02	0.02	17.69	0.07	-	-	0.04	77.31	4.74	-	-	-	-	0.09
[iy]	0.05	0.15	-	-	-	1.64	-	-	0.05	0.45	9.97	87.64	-	-	-	-	0.05
[oh]	1.57	1.44	1.70	-	1.44	1.31	-	-	-	0.13	-	-	**61.52**	14.66	0.39	10.86	4.97
[ow]	0.12	-	1.71	12.93	1.48	0.55	0.04	-	-	-	-	0.08	-	81.19	0.04	0.19	1.67
[oy]	0.10	-	0.10	1.14	-	-	0.41	-	-	0.10	0.10	-	-	2.48	**95.45**	-	0.10
[uh]	-	-	-	-	-	0.36		-	-	-	-	-	-	0.12	-	82.07	17.45
[uw]	-	-	0.38	0.25	0.13	0.19	-	-	0.06	-	0.06	0.31	-	4.75	-	6.38	87.50

Table (5): Consonants Confusion Matrix

Actual Consonants \ Perceived Consonants	[B]	[CH]	[D]	[DH]	[F]	[G]	[HH]	[JH]	[K]	[L]	[M]	[N]	[NG]	[P]	[R]	[S]	[SH]	[T]	[TH]	[V]	[W]	[Y]	[Z]	[ZH]
[B]	63.64	-	-	-	1.12	-	-	-	-	-	-	-	-	35.24	-	-	-	-	-	-	-	-	-	-
[CH]	-	52.70	-	-	-	-	-	0.30	-	-	-	-	-	-	-	0.74	45.77	-	-	-	-	-	-	0.49
[D]	-	-	93.06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6.94	-	-	-	-	-	-
[DH]	-	-	-	66.82	-	-	-	-	-	-	-	-	-	-	-	8.16	-	-	20.02	-	-	-	5.00	-
[F]	-	-	-	-	76.30	-	-	-	-	-	0.72	-	-	0.34	-	-	-	-	-	**22.65**	-	-	-	-
[G]	-	-	-	-	-	83.36	-	-	5.70	-	-	-	10.94	-	-	-	-	-	-	-	-	-	-	-
[HH]	-	-	-	-	-	0.62	99.38	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
[JH]	-	1.90	-	-	-	-	-	79.64	-	-	-	-	-	-	-	0.40	15.37	-	-	-	-	-	-	2.69
[K]	-	-	-	-	-	4.72	-	-	94.35	-	-	-	0.93	-	-	-	-	-	-	-	-	-	-	-
[L]	-	-	-	-	-	-	-	-	-	92.01	-	7.99	-	-	-	-	-	-	-	-	-	-	-	-
[M]	2.42	-	-	-	0.45	-	-	-	-	-	94.98	-	-	1.82	-	-	-	-	-	0.33	-	-	-	-
[N]	-	-	-	-	-	-	-	-	-	1.61	-	81.70	16.69	-	-	-	-	-	-	-	-	-	-	-
[NG]	-	-	-	-	-	5.74	-	-	-	-	-	7.84	86.41	-	-	-	-	-	-	-	-	-	-	-
[P]	27.82	-	-	-	1.27	-	-	-	-	-	-	-	-	70.91	-	-	-	-	-	-	-	-	-	-
[R]	-	-	-	-	-	-	-	-	-	0.26	-	0.66	-	-	97.32	-	-	-	-	-	-	-	-	1.76
[S]	-	0.14	-	0.53	-	-	-	-	-	-	-	-	-	-	-	86.96	0.25	-	9.03	-	-	-	3.09	-
[SH]	-	12.09	-	-	-	-	-	1.14	-	-	-	-	-	-	-	0.41	82.59	-	-	-	-	-	-	3.77
[T]	-	-	0.61	-	-	-	-	-	-	-	-	-	-	-	-	-	-	99.39	-	-	-	-	-	-
[TH]	-	-	-	10.70	-	-	-	-	-	-	-	-	-	-	-	1.00	-	-	83.20	-	-	-	5.10	-
[V]	0.55	-	-	-	31.17	-	-	-	-	-	0.63	-	-	1.10	-	-	-	-	-	66.56	-	-	-	-
[W]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100.00	-	-	-
[Y]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100.00	-	-
[Z]	-	-	-	1.74	-	-	-	-	-	-	-	-	-	-	-	1.14	-	-	3.23	-	-	-	93.79	0.10
[ZH]	-	3.86	-	1.19	-	-	-	9.66	-	-	-	-	-	-	-	1.34	6.39	-	-	-	-	-	-	77.56

4. Experimental Testing

Confidence measure is used in speech recognition for phone error detection algorithms [18]. Therefore, HMM with likelihood score, log posterior probability score, and segment duration score are used to compute phoneme pronunciation score [17].

The user interface was designed using Silverlight technology. This user interface includes different visual properties for basic functions, such as moving between demos, playing a sample (predefined example), testing the user voice, and recording user voice. Figure 5 illustrates the device setting and microphone adjustment.

Figure 5-a: The Device Setting and Microphone adjustment of the Speak Correct System

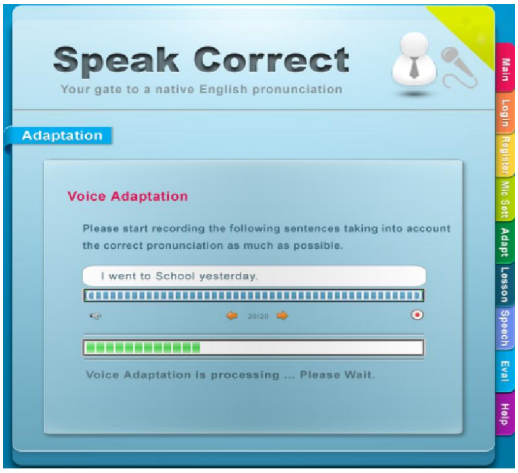

Figure 5-b: The Device Setting and Microphone adjustment of the Speak Correct System

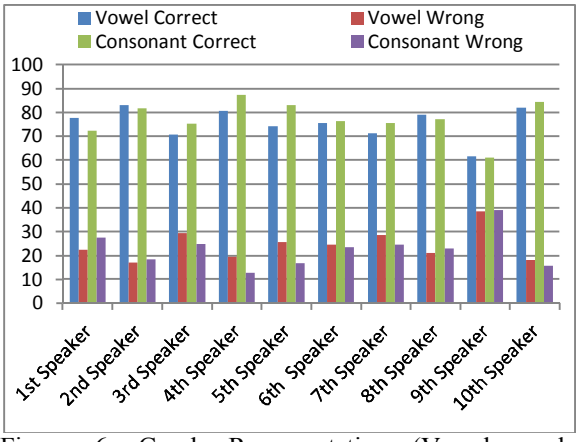

Figure 6: Graph Representation (Vowels and Consonants) of the 10 Students

Pronunciation error detection will be computed, taken into consideration precision-recall for each acoustic model. So, confusion matrix will be composed using correct and wrong pronunciations. For the experiment scoring, each test for the vowels tests; and the consonants test. TP indicates to positive and correctly pronounced and TN indicates to negative and mispronounced. Table 5 displays this confused matrix. FN and FP are false negative and false positive. The recall and precision values are computed using the following formulas:

Recall = TP/ (TP + FP) …….(2)

Precision = TP/ (TP + FN) ……..... (3)

Table (6): Confusion Matrix for 10 Students (Sample Test)

	Vowels		Consonants	
	Correct	Wrong	Correct	Wrong
1st Speaker	77.60	22.40	72.41	27.59
2nd Speaker	83.00	17.00	81.60	18.40
3rd Speaker	70.70	29.30	75.20	24.80
4th Speaker	80.50	19.50	87.22	12.78
5th Speaker	74.26	25.74	83.12	16.88
6th Speaker	75.40	24.60	76.43	23.57
7th Speaker	71.30	28.70	75.40	24.60
8th Speaker	79.00	21.00	77.16	22.84
9th Speaker	61.60	38.40	60.94	39.06
10th Speaker	82.00	18.00	84.26	15.74

Therefore, two human experts are employed to judge the phonemes pronunciation, especially pronunciation variants. The evaluation dataset includes utterances from 10 students with equal males and females. Each student practices using at least 10 examples (pronounced words) from 20 lessons of the *Speak Correct* system. After, each expert discusses the errors and mistakes and decides the correct any transcription errors with tree possibilities:

1. Utterance is correct (accepted by the expert).
2. Utterance is not correct; pronunciation error (reported by all the experts).
3. Human experts disagreed to accept or reject the pronunciation (Not Clear).

Accordingly, *Speak Correct* system decides the value of the confidence score (figure 6), such value is one from two: (1) Correct; and (2) Wrong; pronunciation error or unknown (repeat request).

Tables (7-a & 7-b) illustrate the evaluation results for the *Speak Correct* system relative to the human experts' judgment for the 10 students.

As shown in table 7, for correct speech segments the classification of "Repeat Request" was 8.8% of the total correct words (89.7 %). That is because they had low confidence under the computed threshold, and the system gave a repeat request to avoid the possibility of false alarms.

To evaluate the effect of flow adjustment on system performance, testing dataset is used for models adaptation and run the evaluation on the remaining test set. Table (8) illustrates the system performance with 100, 200, 300 utterances as

adjustment data. The table shows the percentage of correct system feedbacks, which is sum of the highlighted blocks in table (7-b).

Table (7-a): Evaluation Result Relative to Human Experts of the Speak Correct

Speak Correct Judgment	Human Experts' Judgment for each Speaker						
			Correct	Wrong	Precision	Recall	F-Score
	1st Speaker	Correct	77.60	1.4	0.94	0.78	0.85
		Wrong	22.40	4.7			
	2nd Speaker	Correct	67.30	2.0	0.95	0.67	0.79
		Wrong	32.70	3.5			
	3rd Speaker	Correct	73.13	1.6	0.94	0.73	0.82
		Wrong	26.87	4.5			
	4th Speaker	Correct	80.10	1.2	0.95	0.80	0.87
		Wrong	19.90	4.6			
	5th Speaker	Correct	70.22	1.7	0.94	0.70	0.80
		Wrong	29.78	4.4			
	6th Speaker	Correct	68.15	1.9	0.93	0.68	0.79
		Wrong	31.85	5.0			
	7th Speaker	Correct	74.65	1.1	0.94	0.75	0.83
		Wrong	25.35	4.8			
	8th Speaker	Correct	82.14	1.3	0.95	0.82	0.88
		Wrong	17.86	4.1			
	9th Speaker	Correct	79.11	2.1	0.96	0.79	0.87
		Wrong	20.89	3.2			
	10th Speaker	Correct	80.00	1.0	0.95	0.80	0.87
		Wrong	20.00	4.0			
	Total	Correct	75.24	1.53	0.95	0.75	0.84
		Wrong	24.76	4.28			

Table (7-b): Evaluation of *Speak Correct* system relative to experts' judgment

Speak Correct Judgment	Human Experts' Judgment			
	Correct	Wrong	Not Clear	Total
Correct	80.9 %	1.4 %	1.2 %	83.5 %
Wrong	0.0 %	4.7 %	0.2 %	4.9 %
Repeat request	8.8 %	2.1 %	0.7 %	11.6 %
Total	89.7 %	8.2 %	2.1 %	100 %

Table 8: The Progressive Model Adaption's Results of the *SpeakCorrect* System.

Size of Data	**10 Utterances**	**100 Utterances**	**200 Utterances**	**300 Utterances**
Correct Feedback	**84.0%**	**86.6%**	**87.2%**	**87.4%**

As illustrated in table 8, the system performance has improved significantly with additional adaptation improvement in system correct feedbacks. This improvement didn't require much computation load since the models adaptation were performed progressively.

Figure (7) shows the precision recall plot obtained by the experimental testing of *Speak Correct* system.

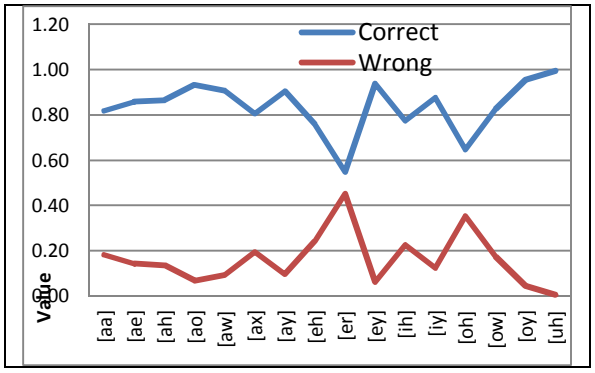

Figure 7: Correct/Wrong Relation for Vowels' Pronunciations

In another way, experimental test for the *Speak Correct* system can be done using word error rate (WER) for each student or speaker.

Conclusion

This paper introduced the *SpeakCorrect* system as Computer Aided Pronunciation Training (CAPT) system for native Arabic students of English pronunciation. Up to now, from the previous discussion, the experimental test is related to unigram features. The bigram and trigram features will be discussed in future work. Elementary evaluation results are promising and show significant improvements in the users' pronunciation skills. The current version of the system only supports phonemic pronunciation errors type. In future work we plan to add practise lessons for the prosodic pronunciation errors.

Acknowledgements

The teamwork of the ***Speak Correct*** project was funded as part of the strategic technology project (10-INF-1406-03) held at the King Abdulaziz University (KAU). Also, the authors wish to thank King Abduaziz City for Science and Technology (KACST) for funding, which was received through grant number 10-INF-1406-03. This financial support during the research period is gratefully acknowledged.

References

1. Shafiro V., E. Levy, R. Dakwer, A. Kharkhurin, Perceptual Confusion of American-English Vowels and Consonants by Native Arabic Bilinguals, Language and Speech Journal, 2012, pp. 1-7.
2. Copper C. and A. Bralow, Perception of dialect Variation in Noise: Intelligibility and Classification, Language and Speech, 2008, Vol. 51, pp. 175-198.
3. Clark M. , Beyond Antagonism? The Discursive Construction of "New" Teachers in the United Arab Emirates, Teaching Education, 2006, Vol. 17, pp. 225-237.
4. Rupp R., Higher Education in the Middle East: Opportunities and Challenges for U.S. Universities and Middle East Partners, Global Media Journal, 2009, Vol. 8, pp. 1-21.
5. Fujii K., T. Yoshioka, K. Yamasaki, M. Muneyasu and M. Morimoto, (2011). A Double Talk Control Method Improving Estimation Speed by Adjusting Required Error Level, Workshop on Hands-free Speech Communication and Microphone Arrays, IEEE May 30 - June 1, 2011.
6. Kac J. and G. Rozinaj, (2009). *Adding Voicing Features Into Speech Recognition Based on HMM in Slovak,* IEEE Conference, Systems, Signals and Image Processing, IWSSIP 2009. 16th International Conference.
7. Abdou S., M. Rashwan, H. Al-Barhamtoshy, K. Jambi, and W. Al-Jedaibi, 2012. *Enhancing the Confidence Measure for an Arabic Pronunciation Verification System.* Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training June 6 - 8, 2012, KTH, Stockholm, Sweden.
8. Jurafsky D. & J. H. Martin, University of Colorado, Boulder, (2008). Speech and Natural Language Processing. 2nd Edition, Prentice Hall.
9. Zhou Y. and L. Shang, (2012). Speaker Recognition Based on Principal Component Analysis and Probabilistic Neural Network, Lecture Notes in Computer Science, 2012, Volume 6839, Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, pp 708-715.
10. Doremalen J., C. Cucchiarini, and H. Strik, Automatic Detection of Vowel Pronunciation Errors Using Multiple Information Sources, Automatic Speech Recognition & Understanding, ASRU 2009, IEEE Workshop, pp. 580-585.
11. Joshi S., P. Rao, Acoustic Model for Pronunciation Assessment of Vowels of Indian English, IEEE COCOSDA Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), International Conference, 2013, pp. 1-6.
12. Gong C., X. Li and X. Wu, The Effect of Part of Speech on Mandarin Speech Recognition, IEEE Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific, 2013, pp. 1-4.
13. Lu L., A. Ghoshal and S. Renals, Acoustic Data-Driven Pronunciation Lexicon for Large Vocabulary Speech Recognition, Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop, Olomouc, 8-12 Dec. 2013, pp. 374 - 379.
14. Wang Y., L. Lee, "Toward Unsupervised Discovery of Pronunciation Error Pattern Using Universal Phoneme Posteriorgram for Computer-Assisted Language Learning", Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference, pp. 8232- 8236.
15. Toma S., T. Birsan, F. Totir, and E. Oancea, On Letter to Sound Conversion for Romanian: a Comparison of Five Algorithms, Speech Technology and Human - Computer Dialogue (SpeD), 7th IEEE International Conference 2013, pp 1-6.

16. Hamadani G., A. Sellouani, M. Boudraa, Effect of Characteristics of Speakers on MSA ASR Performance, 1st IEEE International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 2013, pp. 1-5.
17. Jai J., W. Leung, Y. Tian, L. Cai, and H. Meng, Analysis on Pronunciations in CAPT Based on Computational Speech Perception, Chinese Spoken Language Processing (ISCSLP), 2012 8th IEEE International Symposium, pp. 174- 178.
18. Wang Y., L. Lee, Improved Approaches of Modeling and Detecting Error Patterns with Empirical Analysis for Computer Aided Pronunciation Training, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference, pp. 5049- 5052.
19. Heracleous P.; N. Aboutabit; D. Beautemps; (2009). HMM-based vowel and consonant automatic recognition in Cued Speech for French, Virtual Environments, Human-Computer Interfaces and Measurements Systems. VECIMS '09. IEEE International Conference.
20. Wohlmayr M., M. Stark, and F. Pernkopf, (2011). A Probabilistic Interaction Model for Multi-pitch Tracking With Factorial Hidden Markov Models. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, May 2011.
21. Tsubota Y., T. Kawahara, M. Dantsuji; (2002). Recognition and Verification of English by Japanese Students for Computer Assisted Language Learning System. In Proc. ICSLP, pp. 1205-1208.
22. Witt S. M., Automatic Error Detection in Pronunciation Training: Where we are and where we need to go. Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training June 6 - 8, 2012 KTH, Stockholm, Sweden, (IS ADEPT, Stockholm, Sweden, June 6-8 2012).
23. Pellom B., Rosetta Stone ReFLEX: Toward Improving English Conversational Fluency in Asia. Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training June 6 - 8, 2012 KTH, Stockholm, Sweden, (IS ADEPT, Stockholm, Sweden, June 6-8 2012).

5/23/2014