# Speak Correct: A Computer Aided Pronunciation Training System for Native Arabic Learners of English

[1]Sherif Abdou, [2]Mohsen Rashwan, [3]Hassanin Al-Barhamtoshy, [3]Kamal Jambi and [3]Wajdi Al-Jedaibi
[1]Faculty of Computers & Information, Cairo University, Egypt [2]Faculty of Engineering, Cairo University, Egypt
[3]Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia
(sheriff.abdou ,mrashwan)@rdi-eg.com, (hassanin , kjambi, waljedaibi)@kau.edu.sa

**Abstract**. In this paper we introduce the SpeakCorrect system which is a Computer Aided Pronunciation Training (CAPT) system for native Arabic students of English. The system is designed with optimized performance for the target users group. It is L1 dependent system and only the frequent pronunciation errors of native Arabic speakers are examined. Several adaptation techniques such as Speaker Adaptive Training (SAT), Speaker Clustering (SC) and Maximum Likelihood Linear Regression (MLLR) are used to boost the performance of the SpeakCorrect system. The decision reached by the SpeakCorrect system is accompanied by a posterior based confidence score to reduce effect of misleading system feedback.Evaluation results for the system are promising and show significant improvements in the users' pronunciation proficiency.
[Sherif Abdou, Mohsen Rashwan, Hassanin Al-Barhamtoshy, Kamal Jambi and Wajdi Al-Jedaibi. **Speak Correct: A Computer Aided Pronunciation Training System for Native Arabic Learners of English.** *Life Sci. J* 2014; 11(10):370-380] (ISSN: 1097-8135). http://www.lifesciencesite.com. 51

**Keywords:** Computer Aided Pronunciation Training, Speaker Adaptation, Confidence Measure

## 1. Introduction

With increasing globalization, there has also been a significant increase in the demand for foreign language learning. One aspect of which is pronunciation learning. The need to speak a foreign language that is different from the mother tongue (such as English vs. Arabic) may lead to severe pronunciation problems, due to the difficulty of hearing the differences between the own pronunciation and what is correct. “A foreign language is not only a question of getting the words and syntax right. You can’t be understood until you can pronounce it well. Effectively teaching pronunciation typically requires one-to-one teacher student interactions, which for many students is unaffordable. For this reason, automatic pronunciation teaching has been a focus of the research community(Silke 2012).

With the increased computing power and the rapid progress in computer-based speech processing along with the creation of advanced methods for speech recognition(including dialects and accents) now make it possible to apply modern speech technologies to "Computer-Aided Pronunciation Teaching' (CAPT) (Delmonte, 2011).There are currently available several systems that can measure the pronunciation quality of students by analysing few minutes of their speech and have shown to be as reliable as trained human experts (Bernstein, 2010). While this high-level global pronunciation scores might be sufficient for oral proficiency and pronunciation assessment purposes but in general it is not detailed enough for training purposes (Olov, 2012).

For pronunciation training the student must identify on which phoneme the error occurred, diagnose in what way his production differed from the model and understand how this could be corrected (Abdou *et al.,* 2012). For that purpose a pronunciation error detection is required, that is the procedure by which a score at a local (e.g. phoneme) level is calculated. Several approaches have been proposed for phone level pronunciation error detection. Most of these approaches use Automatic Speech Recognition (ASR) based metrics such as log-likelihood scores, posterior probabilities and (log) likelihood ratios (Kim, 1997, Silke, 1999, Franco, 2000).The later one has become a de-facto standard for judging the goodness of phone pronunciation, since it was shown that it had the highest correlation with human scores (Silke, 2012).

One of the core decision points for pronunciation error detection is whether to build a system that is L1 (i.e. the native language) dependent or not. Better performance has been found with methods that take L1 into account. This approach has two main advantages: Firstly, if L1 is known, one can utilize acoustic models that are a mixture of L1 and L2 (Hui, 2005, Saz, 2009) and have improved speech recognition accuracy, which in turn enables recognition of less constrained utterances, which allows for greater freedom in the selection of pronunciation learning exercises, in particular for assessing fluency. Secondly, the set of common pronunciation errors tend to be typical for a given L1 and very different between different L1, i.e. a native Arabic speaker will make very different English pronunciation errors than a native speaker of French

or Chinese. Thus, knowledge of L1 enables to provide tailored pronunciation exercises (Johnson 2012).

Also it has long been known that the differences between native and non-native speech are so extensive as to degrade ASR performance considerably (Doremalen, 2010, 2011). To reduce this effect some CAPT systems adapt its acoustic models to match the characteristics of the user voice. It was shown that using standard adaptation algorithms such as MAP or MLLR yields substantial recognition accuracy improvements and going from speaker independent to speaker dependent recognition almost reduces the phoneme recognition error rate in half (Hui, 2005, Saz, 2009).

In this paper we introduce the Speak Correct system which is a CAPT system that is designed to teach English pronunciation for Native Arabic speakers (Al-Barhamtoshy, 2014). The system is designed with targeted optimum performance. It uses a state of art speech recognizer. The system is L1 dependent and only the frequent pronunciation errors of native Arabic speakers are examined by the speech decoder. Several adaptation techniques are used to boost the performance of the SpeakCorrect system. Initially during the training phase the Speaker Adaptive Training (SAT) (Anastasakos 1996) technique is used to reduce the inter-speakers variability, including the non-native speech effects, in the training data. In the testing phase a speaker classification is used to map the speaker to the nearest cluster of speakers which their data are used to adapt the system models. Finally the system models are adapted by the Maximum Likelihood Linear Regression (MLLR) adaptation using few words of the user. Whenever more data is available from the user a cascade adaptation technique is used to keep enhancing the system performance. The decision reached by the Speak Correct system is accompanied by a posterior based confidence score to reduce effect of misleading system feedback.

In the following sections, section 2 summarizes the phonetic language differences between Arabic and English and common pronunciation error patterns. The SpeakCorrect CAPT system is described in section 3. Section 4 describes the multiple adaptation techniques utilized in the system. Section 5 describes the used confidence score. Section 6 describes the system user interface. Section 7 includes some evaluation results for the system and section 8 includes the final conclusions and planned future work.

## 2. Common Pronunciation Error Patterns of Arabic Learners of English

Before developing the Speak Correct system, we obtained an overview of frequent errors made by native Arabic language learners of English. Our sources of information were literature information, expertise of language teachers, and analysis of a collected database from 200 students.

The Arabic and English phonological systems are very different, not only in the range of sounds used, but in the emphasis placed on vowels and consonants in expressing meaning. While English has 22 vowels and diphthongs and 24 consonants, Arabic has only eight vowels and diphthongs (three short, three long and two diphthongs) and 32 consonants.

The three short vowels in Arabic have very little significance: they are almost allophonic. They are not even written in the script. It is the consonants and long vowels and diphthongs which give meaning. Arabic speakers tend, therefore, to gloss over and confuse English short vowel sounds, while unduly emphasizing consonants, avoiding elisions and shortened forms.

Among the features of Arabic which give rise to an 'Arabic accent' in English are:

- More energetic articulation than English, with more stressed syllables, but fewer clearly articulated vowels, giving a dull, staccato 'jabber' effect.
- The use of glottal stops before initial vowels, a common feature of Arabic, thus breaking up the natural catenations of English.
- A general reluctance to omit consonants, once the written form is known, e.g./klaimbed/ for climbed.

**Vowels**

Table 1: The English and Arabic vowels map (using IPA symbols)

| | | | | | |
|---|---|---|---|---|---|
| ʌ | | Cup | ɔ: | | four |
| æ | ـَ | Cat | u: | و | food |
| ɑ: | ـَ | father | əʊ | | home |
| e | | Met | aɪ | | five |
| ə | | cinema | aʊ | ـُ | now |
| ɜ: | | Learn | eɪ | ـِ | raid |
| ɪ | ـِ | Hit | ɔɪ | | boy |
| i: | ي | Heat | eəʳ | | where |
| ɒ | | Hot | ɪəʳ | | near |
| ʊ | ـُ | Put | ʊəʳ | | pure |

Shaded phonemes in table (1) have equivalents or near equivalents in Arabic and should therefore be perceived and articulated without great difficulty,

although some confusion may still arise. The Modern Standard Arabic (MSA) includes equivalent 6 phones is:

æ→ َ (fatha)
ɑ:→ ً (Emphatic fatha)
ɪ→ ِ (kasra)
i: → ي (long kasra)
ʊ→ ُ (dama)
u:→ و (long dama)

The Egyptian Colloquial Arabic includes 2 extra phonemes:

eɪ→ ِ (tilted kasra)
aʊ→ ُ (tilted dama)

Unshaded phonemes in table (1) may cause problems. While virtually all vowels may cause problems, the following are the most common confusions:

(1) /ʌ/ "cup" vs. /æ/ "cap"
(2) /ɪ/ "sit" vs. /e/ "set"
(3) /ɒ/ "cot" vs. /ɔ:/ "caught"
(4) Diphthong /eɪ/ /əʊ/ are usually pronounced rather short and is confused with their equivalent short vowels:
   (a) /e/ "red" for /eɪ/ "raid";
   (b) /ɒ/ "hop" for /əʊ/ "hope

**Consonants**

Table 2: The English and Arabic consonants map (using IPA symbols)

| | | | | | |
|---|---|---|---|---|---|
| P | | Pin | l | ل | sell |
| b | ب | Bin | tʃ | تش | church |
| m | م | Man | ʒ | ج | jar |
| f | ف | Fan | ʃ | ش | shin |
| v | | Van | dʒ | | leisure |
| θ | ث | Thin | r | ر | narrow |
| ð | ذ | Bathe | j | ي | year |
| t | ت | talked | k | ك | kin |
| d | د | Lid | g | ج | get |
| s | س | Sin | ŋ | | sing |
| z | ز | Zoo | w | و | which |
| n | ن | Pin | h | ه | hat |

Shaded phonemes in table (2) have equivalents or near equivalents in Arabic and should therefore be perceived and articulated without great difficulty, although some confusions may still arise. Unshaved phonemes may cause problems. The following are the most common confusions:

(5) Arabic has only one letter in the /g/—/dʒ/ area, which is pronounced as /g/ in some regions, notably Egypt, and as /dʒ/ in others. Arabic speakers tend, therefore, to pronounce an English g, and sometimes even a j, in all positions according to their local dialects.
(6) /g/ "garden" → /ʒ/ "jarden"
(7) /ʒ/ "jury" → /g/ "gury"
(8) /tʃ/ as a phoneme is found only in a few local Arabic dialects, but the sound occurs naturally in all dialects as junctures of /t/ and /ʃ/. But it is very common error to miss the /t/ sound.
/tʃ/ "church" → /ʃ/ "shurch "
(9) There are two approximations to the English /h/ in Arabic "ه" and "ح". The first of them which is an unvoiced harsh aspiration is more commonly used for the /h/ sound. So Arabic speakers tend therefore to pronounce an English /h/ rather harshly.
(10) Arabic speakers tend to speak the /r/ phoneme with a rhotic accent and pronounce it as a flap or trill. Clearly this pronunciation error results from the effect of the equivalent Arabic phoneme, the /R/ "ر"sound. Arabic speakers commonly over pronounce the post-vocalic /r/, as in "car" "park".
(11) The bilabial plosives /p/ and /b/ are allophonic and tend to be used rather randomly:
(12) I baid ten bence for a bicture of Pig Pen.
(13) The labio-dental fricative /v/ sound does not exist in Arabic so it frequent to be confused with its allophonic sound /f/:
(14) It is a fery nice fillage.
(15) /g/ and /k/ are often confused, especially by those Arabs whose dialects do not include the phoneme /g/. These pairs usually cause difficulty:
(16) /g/ "goat" Vs. /k/ "coat"
(17) /g/ "bag" Vs. /k/ "back"
(18) Although /θ/ and /ð/ occur in literary Arabic, most dialects pronounce them as /t/ and /d/ respectively. The same tends to happen in students' English.
(19) I tinkdatdey are brudders.
(20) Sometimes they are confused with each other or even with the /s/ and /z/ sound as:
   (a) /θ/ "thank" → /s/"sank"
   (b) /ð/ "father" → /z/ "fazer"
   (c) /z/ "prize" → /s/ "price"
(21) The phoneme /n/ is usually pronounced as /n/ or /ng /, or even /nk/.
   "morn**in**" → "morn**ing**" or "morn**ink**"

**Consonant Clusters**

The range of consonant clusters occurring in

English is much wider than in Arabic. Initial two-segment clusters not occurring in Arabic include: “pr”, “pl”, “gr”, “gl”, “thr”, “thw”, “sp”. Initial three-segment clusters do not occur in Arabic at all, e.g.: “spr”, “skr”, “str”, “spl”. In all of the above cases there is a tendency among Arabic speakers to insert short vowels to 'assist' pronunciation:

“price” → “p**e**rice” or ”"p**i**rice”
Spring → “**i**spring” or “s**i**pring”

The range of final clusters is also much smaller in Arabic. Of the 78 three-segment clusters and fourteen four-segment clusters occurring finally in English, none occurs in Arabic. Arabic speakers tend again to insert short vowels (Nasr 1963).

“arranged” → “arrang**i**d”
“months” → “month**i**z”
“next” → “"neck**i**st”

**Influence of English Spelling on Pronunciation**

While there are no similarities between the Arabic and English writing systems, Arabic spelling within its own system is simple and virtually phonetic. Arabic speakers tend, therefore, to attempt to pronounce English words phonetically. Add to this the reverence for consonants, and you get severe pronunciation problems caused by the influence of the written form:

“stopped” → “"istobb**i**d”
“foreign” → “for**ige**n”

**Juncture**

As the glottal stop is a common phoneme in Arabic, and no words begin with a vowel, there is resistance in speaking English to linking a final consonant with a following initial vowel. Junctures producing consonant clusters will cause problems, as described under the section 'Consonant clusters'. A juncture such as "next spring" produces a number of extra vowels. Also the many instances of phonetic change in English through the juncture of certain phonemes,

e.g.

/t/ + /j/ as in “What you need” → /wDtʃu:ni:d/,

/d/ + /j/ as in “Did you see him? → /dɪdʒu:si:hɪm/

/n/ changes to /m/ before /m/ /b/ or /p/ as in “can be”→ “ca**m**bi:”

These changes are resisted strongly by Arabic speakers, who see any loss of or change in consonant pronunciation as a serious threat to communication but if they don't use them their speech would sound too formal.

**3. The SpeakCorrect Tool Architecture**

Figure 1 shows a block diagram of the Speak Correctsystem. It uses a state of art speech recognizer with Hidden Markov Models (HMM) to detect pronunciation errors in the speech of the users. Its main blocks are:

- **The HMM models trainer**: Collect statistical patterns from the training data and save them in statistical models that are used in the pronunciation verification phase.
- **Verification HMM models**: Is the acoustic HMM models for the system.
- **Pronunciation hypotheses generator**: It analyzes a training exercise and generates all possible pronunciation variants that are fed to the speech recognizer in order to test them against a spoken utterance.
- **The HMM Adapter**: Is used to adapt acoustic models to each user acoustic properties in order to boost system performance.
- **The HMM Decoder (ASR)**: The decoder that recognizes the user input speech.
- **Confidence Measure**: It receives n-best decoded word sequence from the decoder, then analyzes their scores to determine whether to report that result or not.
- **The Pronunciation Errors Analyzer**: Analyzes results from the speech recognizer and produce the feedback messages to the user.
- **The Intonation analyzer**: Analyzes the pitch curves for the user utterance and gives feedback messages for the prosodic and rhythm errors.
- **Feedback generator**: map the detected errors to feedback messages that explain to the user his faults and guide him to enhance his pronunciation.

The tool performs two main tasks. Firstly it recognizes a mispronounced utterance, even if it is pronounced in a deviant way; and secondly, it locates at the phoneme level the pronunciation errors made by the speaker. These two tasks are implemented in two modules in the system, the Automatic Speech Recognizer (ASR) and the Pronunciation Analyser (PA). The role of the ASR is to transcribe the user's utterances to the system, while the pronunciation analyser uses the output from the ASR to judge whether the pronunciation is accepted as correct or not and to spot prototypically deviant phonemes (i.e. finding on what part of the utterance the feedback should be focused).

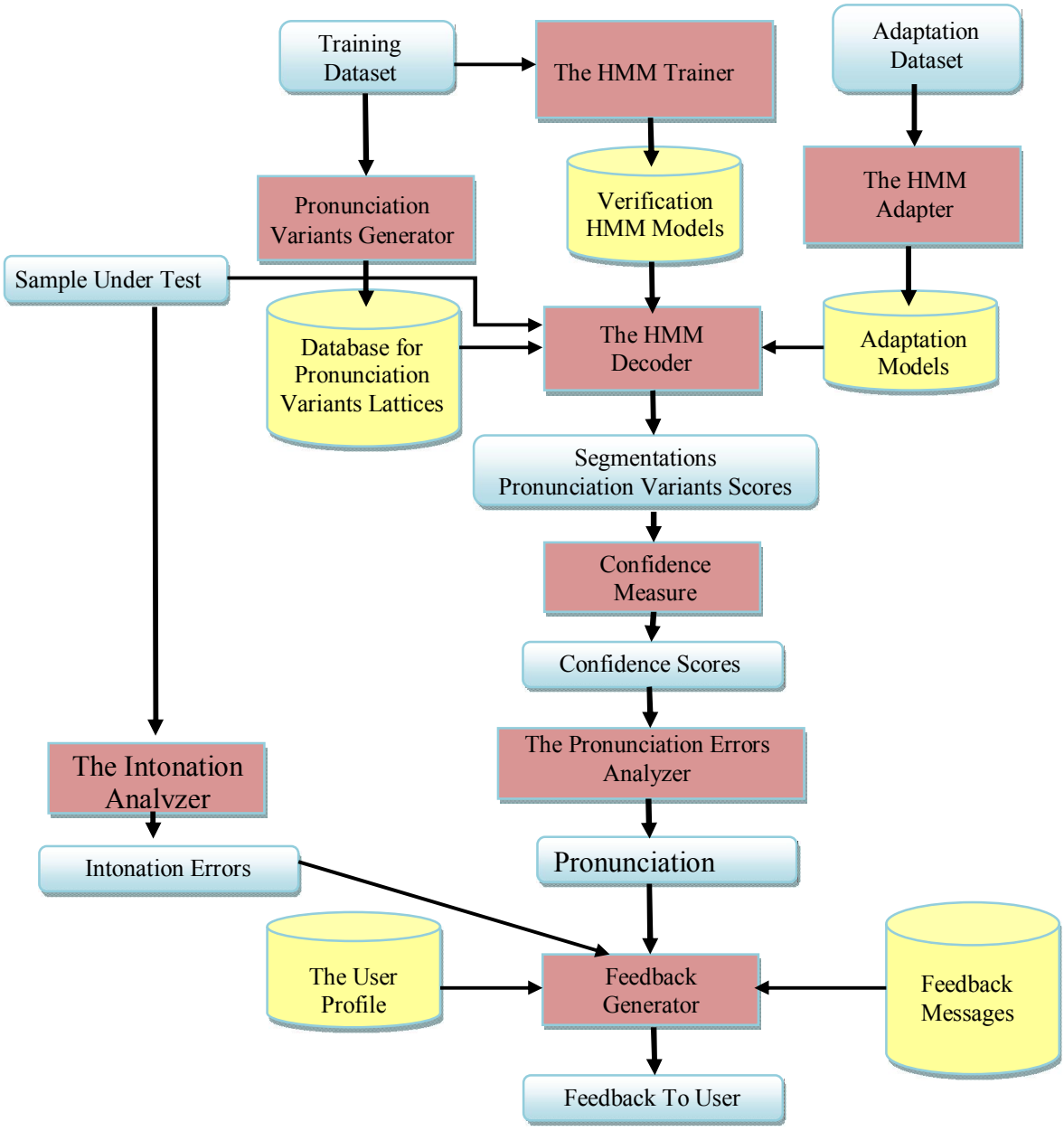


**Figure 1:** The SpeakCorrect System Architecture

To increase the accuracy of the system, only probable pronunciation variants, that cover common types of pronunciation errors, are examined. An approach for automatic generation of pronunciation hypotheses is used (Hamid, 2005). In that approach the pronunciation hypotheses are reached by deploying matching rules to detect pronunciation patterns and generate corresponding probable pronunciation errors. The following two sections describe the details of the Speak Correct models adapter and the confidence scoring module.

**4. The Speak Correct System Models Adapter**

During the usage of the tool a speech data are accumulated for every user. This data can be used to adapt the generic, speaker independent, acoustic models so it becomes closer to the acoustic prosperities of a specific user. This models adaptation results in significant improvement in the accuracy of the tool.

The goal is to adapt the reference acoustic models parameters to the user acoustic properties. This is done in three steps, first the system user is

classified to the nearest speaker cluster. Secondly, Maximum Likelihood Linear Regression (MLLR) speaker adaptation algorithm is used (Leggetter, 1996). Finally, supervised incremental adaptation is used to refine the adapted model.

So when using speaker adaptation in CAPL systems it is desired to model the speaker special acoustic features that are due to his/her gender, age and physical properties of his/her speech production system. Therefore, models should not be affected by the speaker's special accent and certainly not by his common mispronunciations. The problem can be stated in another form: How to use data of a new speaker that the system has no assumption of the quality of his pronunciation to adapt a carefully built reference acoustic models set with low variance in order to be able to detect mispronunciations?

**The Models Adaptation Procedure**

To meet the requirements of speaker adaptation in the SpeakCorrect system the following algorithm is used:

- **Step 1:** First collect few common sentences from the speaker so as to assign the user to a certain cluster of speakers (Kosaka, 1994). Each user uses his cluster's transformation during the data collection part of the enrolment process. This reduces the gap between reference models and the new speaker characteristics which speed up collection of adaptation data.
- **Step 2:** Prompt the user to utter phrases and test them with adapted reference models generated in step 1. If the system decides that an utterance is accepted (the acceptance criteria is described in the following section), add it to the group that is used in speaker adaptations.
- **Step 3:** Continue until the amount of collected adaptation data is sufficient to apply MLLR speaker adaptation technique to transform reference models to the current speaker's domain.
- **Step 4:** As the system is collecting user data, the cascade adaptation mechanism (described in the following section) is used to enhance the user profile.

Figure 2 shows a block diagram of the adaptation process in the SpeakCorrect system, which shows the process steps and data flow diagram throughout it.

**Usability issues in the User Enrolment Process**

During the first step of the user enrolment process, it is required to collect few utterances (two was selected in our system) to initially adapt the reference model. In this phase it is necessary that the utterances used are carefully selected to be sure that the user will most probable pronounce them accurately. Common utterances are preferred candidates, as by this way probability of mispronunciation is minimized. Also, it is desirable to inform the user that the system will not check pronunciation and that pronouncing them inaccurately will greatly slow adaptation collection and may affect overall system performance (Al-Barhamtoshy *et al.,* 2014).

The acceptance criteria of utterances in the second step are crucial because the adaptation data collection process should be reliable and should not take more than few minutes (in our case, less than 10 minutes). So rather than using the correctness of utterances as the only criteria for accepting an utterance, we decided to involve confidence scoring. A confidence threshold is imposed on each decoded phone, while another threshold is put on the number of wrongfully uttered phones that is accepted in the system by confidence. So even if the user uttered the sentence wrongfully, it will be used in the adaptation process using the decoded sequence rather than the expected correct phone sequence, if we have high confidence in the decoder output.

**Cascade Adaptation**

As users practice pronunciation using the SpeakCorrect system, users' profiles are being enhanced using the collected data for each user because speakers' transformations are getting more accurate. One option to create the new transformation for a user is to use all data collected from that user. As the accumulated data in the user profile gets larger this process will be time consuming. Alternatively, for the SpeakCorrect system we developed a cascade adaptation approach that can reduce the processing time required to create the new speaker transformation with minimal degradation in performance. The idea of that approach is to calculate the new transformation based on the old transformation with the newly collected data only (assuming that old data is represented in the old transformation). But in this case, more than one transformation should be applied every time the decoder is initiated (as the number of transformations to be applied will increase linearly). This problem is solved by using transformation summation and a recursive tree search as explained in (Samir, 2007).

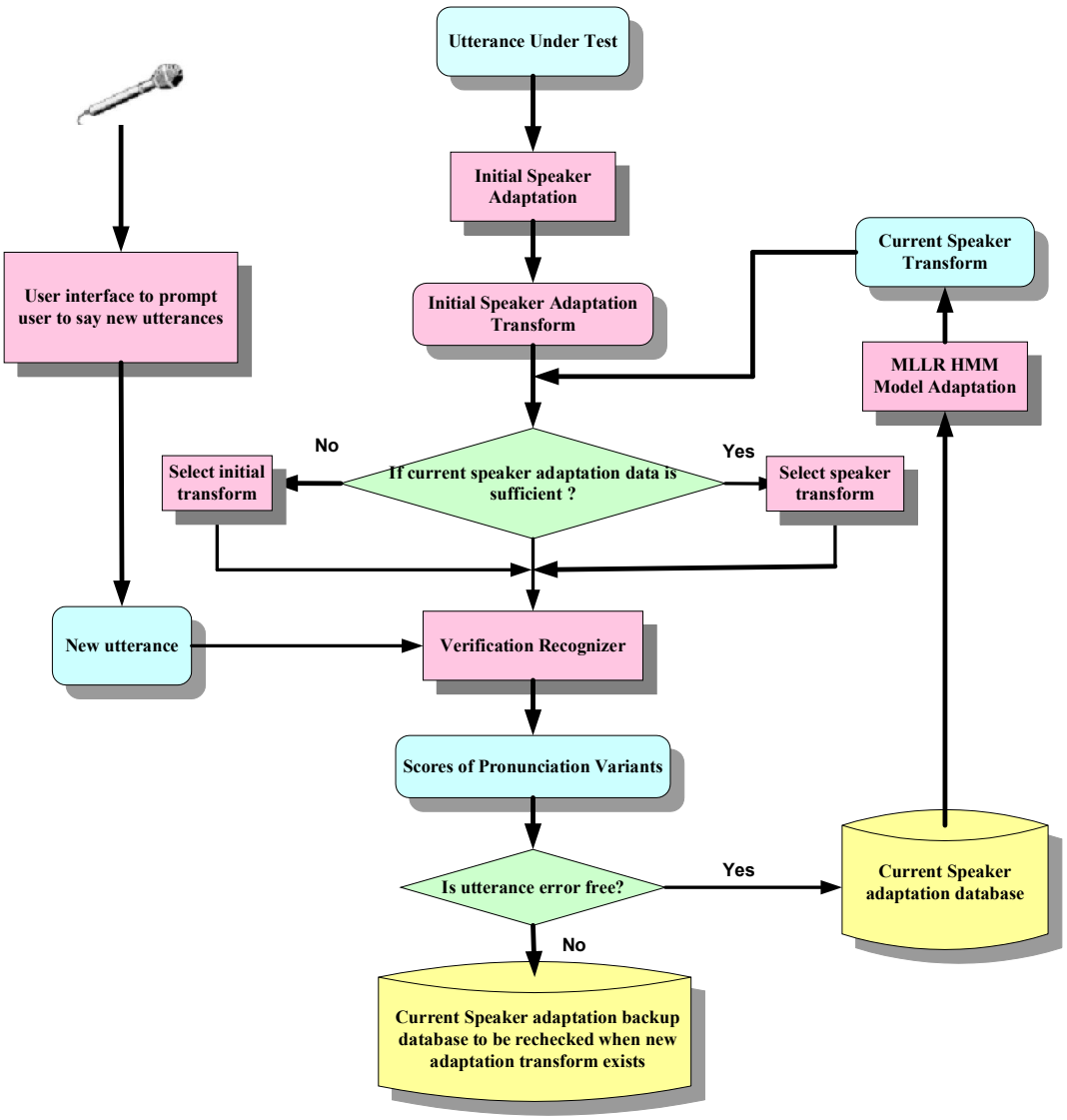


Figure.2: Block Diagram for the Speaker Adaptation Process in Speak Correct System

**5. Confidence Measure**

Every pronunciation error analysis generated by the tool will be associated with a corresponding confidence score that is used to choose suitable feedback response to the learner. When the system suspects the presence of a pronunciation error with low confidence score the system will have those alternate responses:-

(1) Omit the reporting of the error at all (which is good for novice users because reporting false alarms discourages them to continue learning correct pronunciation).
(2) Ask the user to repeat the utterance because it was not pronounced clearly.
(3) Report the existence of an unidentified error and ask the user to repeat the utterance (which is better for more advanced users than ignoring an existent error or reporting wrong type of pronunciation error).
(4) Report most probable pronunciation error (which if wrong- can be very annoying to many users).

In order to reduce effect of misleading system feedback to unpredictable speech inputs, the decision reached by the recognizer in the SpeakCorrect system is accompanied by a confidence score. The implemented confidence scoring in SpeakCorrect is based on the Likelihood ratios(Williams, 1999)where the acoustic model likelihoods are scaled by the

likelihood of the first alternative path model as the competing decode model. During decoding process, the Viterbi decoder at the end of each decoded sub-word $M_{Best}$– at frame $x_E$- backtracks in the recognition lattice at both the decoded path and the first alternative path $M_{1st_alt}$until it reaches the node where the two paths meet at the same frame $x_S$. Then it calculates the average confidence score per frame using the formula:

$$CS = \frac{1}{N}\sum_{i=S}^{E}\frac{P(x_i \mid M_{best})}{P(x_i \mid M_{1st\_alt})}$$

Where, N is the number of frames, N = E – S.

Because the difference between these two paths may be significant only in small portion of the path, these small portions should have the most significant effect on the computed confidence score. Therefore, the confidence score of each path is weighted by the distance between the two competing models estimated using Euclidian distance between the center of gravity of the two probability distributions(Hamid, 2005).

## 6. The Speak Correct System User Interface

When a new user registers for the system, he enters his basic information such as gender, age and nationality. This information is used to select the best model that match the user. Then the user passes through the enrolment process to collect sample utterances, with total duration around 2 minutes, from his voice for the initial models adaptation. Figure (3) shows the application screens for this step.

After that stage the user can start practicing with the pronunciation exercises. He can select the lesson he wants to work on. The user starts the recording process of his speech. The system automatically detects his end of speech with the "Silence Detector" and stop the recording automatically. Then after judging and evaluating the user pronunciation a feedback is given to him. Either with green mark that his pronunciation is perfect or red marks for errors. In case of errors, a message is displayed to the user, and also played as audio, that he has committed an error with the error type and guiding instructions to help the student reduce his accent effect. The exercise screens are shown in Figure (4).


**Figure (3):** The Device Setting and Voice Adaptation for the Speak Correct System


**Figure (4):** The Pronunciation Practicing Lessons of the Speak Correct System

**7. System Evaluations**

As the system target is to be used by learners, so the best system evaluation is the one based on the degree of user benefit from the system responses. We developed a new automatic evaluation technique. This technique evaluates the system by measuring the degree of usefulness of its feedback to learners. Evaluating a CAPL system by this means emphasize the system responses for confident decisions and make general feedbacks, or no comments for non-confident decisions to reduce deceiving effect of inherent speech recognition systems limited accuracy. Automation of the evaluation process is vital due to complexity of CAPL systems and the existence for many tuneable thresholds and parameters.

Human experts sometimes disagree on one judgment on a phoneme pronunciation. There is no sharp boundary separating the pronunciation variants, and pronounced sound sometimes lies between two probable pronunciation variants. Also over concentration on a fatal pronunciation mistake can make an expert disregardan adjacent minor mistake. Though we found this disagreement is less than 3% of the evaluation database, when the system approached high accuracy decisions, this disagreement percentage constitutes a considerable amount of noise added to the system evaluation. Also confidence measures used in the system enables the system to generate general and/or ambiguous feedbacks to the student that can't be directly compared to human experts' hard-decision transcriptions.

The evaluation database contains utterances from40 users. Those users are native Arabic male and female college students in different grades. Each student was requested to practice with the SpeakCorrect system by trying at least 10 examples from each one of the 30 lessons of the system. The examples were randomly selected to confirm the inclusion of the whole set of the system examples in the testing database. The total dataset is 12756 utterances. Some utterances were excluded as they were un-complete trails. These utterances was evaluated by a number of language experts, and labelled with the actual pronounced phonemes. Each expert was allowed to transcribe the utterances in a separate session to avoid the possibility that his decision is affected by his colleagues' opinions. For ambiguous speech segments experts were allowed to write all acceptable judgments in their opinions. After each expert has finished, all experts' transcriptions are summed to produce a list of all the judgments accepted by the experts. Afterwards, a final group session is held where all experts discuss each error and they can agree on either to keep all the judgments or choose one or more of them, that's to correct any transcription errors that may be generated by them.

The database is splatted into two parts:

- Calibration set: for calculating optimum values for system parameters and calibrating confidence score thresholds. This set included 6000 utterances.
- Evaluation set: used for the final evaluation of the system. This set included 6250 utterances.

For the evaluation database, the judgment has three possibilities:

(1) Correct (accepted by all human experts).
(2) Identified pronunciation error (all human experts reported the same type of error).
(3) Not Perfect (human experts disagreed whether to reject or accept the pronunciation). That can happen when the pronunciation of a segment is not perfectly correct.

For system judgments, the system keeps track of the best two alternative pronunciations for each speech segment and then computes the confidence score. The state of the best two alternatives is one of three states:

(1) The best alternative is the correct pronunciation, and the second alternative is a pronunciation error.
(2) The first alternative is a pronunciation error, and the second is the correct pronunciation.
(3) The first two alternatives are pronunciation errors.

For each of the previous cases we define a threshold that separates high and low confidence. If the confidence score is above the threshold, the system reports correct pronunciation for the first case, or pronunciation error with the type of error according to the best scoring alternative for the other two cases.

If the confidence score is below the threshold, the system considers the judgment unidentified and the system asks the user to repeat the example. Except for the third case, because the first two alternatives are errors, so we assume the user mispronounced the specified phoneme although the system is not sure of the type of the error. Table (3) shows the evaluation results for the Speak Correct system. Therefore, the system judgment is one of four:

(1) Correct
(2) Pronunciation error with the specified error type
(3) Unknown whether correct or wrong(repeat request)
(4) Error with an unidentified error type

As we see in table (3), for correct speech segments the system yielded "Repeat Request" for about 9.7% of the total correct words. That is because they had low confidence below the computed threshold, and the system gave a repeat request to avoid the possibility of false alarms.

For wrong speech segments, which constitute 8.2% of the data, the system correctly identified the error in 50% of pronunciation errors, reported unidentified errors for 4.8% and gave "Repeat Request" for 25.6% of the errors. The system made false acceptance of 17% of total errors.

The results in table (3) are for the system users after passing the basic adaptation step as described before. To evaluate the effect of cascade adaption on system performance, we used some of the testing dataset for models adaptation and run the evaluation on the remaining test set. Table (4) shows the system performance after using 100, 200, 300 utterances as adaption data. The table show the percentage of correct system feedbacks, which is sum of the highlighted blocks in table (3).

**Table 3:** SpeakCorrect System Evaluation Results for 20 Random Users

| | | Human Judgment | | | |
|---|---|---|---|---|---|
| System Judgment | | Correct | Wrong | Not Clear | Total |
| | Correct | 80.9% | 1.4% | 1.2% | 83.5% |
| | Wrong With Same Error Type | 0 | 4.1% | 0.2% | 4.5% |
| | Wrong With Wrong Error Type | | 0.2% | | |
| | Repeat Request | 8.8% | 2.1% | 0.7% | 11.6% |
| | Wrong With Unidentified Error | 0 | 0.4% | 0% | 0.4% |
| | Total | 89.7% | 8.2% | 2.1% | 100% |

**Table 4:** The Results for the Progressive Models Adaptation for the Speak Correct System

| **Size of Adaptation Data** | **100 Utterances** | **200 Utterances** | **300 Utterances** |
|---|---|---|---|
| **% Correct Feedback** | 86.6% | 87.2% | 87.4% |

From results in table (4), we can see that the system performance has improved significantly with additional adaptation with absolute 2.4% improvement in system correct feedbacks. This system improvement did not require much computation load since the models adaptation were performed progressively.

## 8. Conclusions

In this paper we introduced the SpeakCorrect system which is a Computer Aided Pronunciation Training (CAPT) system for native Arabic students of English. The system is designed with target optimized performance for the target users group by limiting the search space for the set of frequent errors, using posterior based confidence scores and adapting the system models to match the characteristics of the user voice. 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 our future work, we plan to add practise lessons for the prosodic pronunciation errors.

## Acknowledgment

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

Silke M. Witt, (2012). 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).

Olov Engwall, (2012)"Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher", Computer Assisted Language Learning, Vol 25, No. 1, p. 37-64, February 2012.

Lewis Johnson. (2012) "Error Detection for Teaching Communicative Competence", ISADEPT 2012, Stockholm, Sweden, June 2012.

Abdou S., M. Rashwan, Hassanin Al-Barhamtoshy, Kamal Jambi, and Wajdi 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. www.speech.kth.se/isadept/programme.php.

Al-Barhamtoshy H., Abdou S., Jambi K.. (2014) "Pronunciation Evaluation Model for None Native English Speakers", Life Science Journal, 11 (9), http://www.lifesciencesite.com.

Rodolfo Delmonte. (2011) "Exploring Speech Technologies for Language Learning", http://www.intechopen.com/books/speech-and-languagetechnologies,

Jared Bernstein, Alistarit Van Moere, Jian Cheng.(2010) "Validating automated speaking tests", Language Testing, Vol 27, p 355-377.

Silke M. Witt. (1999) "Use of Speech recognition in computer-assisted language learning", unpublished thesis, Cambrige Uni. Eng. Dept.

Yoon Kim, Horacio Franco, and Leonardo Neumeyer. (1997) "Automatic pronunciation scoring of specific phone segments for language instruction", Eurospeech, Rhodes, Greece.

Franco H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000) Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30, 121-130.

Hui Ye, Steve Young. (2005) "Improving the Speech Recognition Performance of Beginners in Spoken Conversational Interaction for Language Learning", Interspeech, Lisboa, Portugal

Oscar Saz, Eduardo Lleida, and William Rodríguez. (2009) "Acousticphonetic decoding for assessment of mispronunciations in speakers with cognitive disorders", AVFA09, 2009.

van Doremalen J., C. Cucchiarini, H. Strik (2010) Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP Journal on Audio, Speech, and Music Processing, Article ID 973954, 13 pages.

Van Doremalen J., C. Cucchiarini, H. Strik (2011) Speech Technology in CALL: The Essential Role of Adaptation. Interdisciplinary approaches to adaptive learning; Communications in Computer and Information Science series, Volume 26, pp. 56-69.

Anastasakos T., J. McDonough, R. Schwartz, J. Makhoul, (1996) " A Compact Model for Speaker-Adaptive Training " Proceedings ICSLP October, Philadelphia, PA

Bernard S. (2011). Arabic Speakers: Learner English, Cambridge Handbooks for Language Teachers, 2nd Edition, Series Editor Scott Thornbury.

Raja T. Nasr , (1963) "Teaching of English to Arab Students" by (Longman, 1963)

Hamid S. (2005). "Computer Aided Pronunciation Learning System using Statistical Based Automatic Speech Recognition. PhD thesis, Cairo University, Cairo, Egypt.

Kosaka T., and S. Sagayama, (1994). "Tree–structured speaker clustering for fast speaker adaptation", proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 1, 245–248. IEEE, New York.

Samir A., S. M. Abdou, A. H. Khalil, M. Rashwan, (2007)."Enhancing usability of CAPL system for Qur'an recitation learning", INTERSPEECH - ICSLP, Antwerp, Belgium.

Williams D. A. (1999). "Knowing what you don't know: roles for confidence measures in automatic speech recognition", Ph.D. thesis, Department of Computer Sciences, University of Sheffield, Sheffield, United Kingdom.