

# Performance Tuning and System Evaluation for Computer Aided Pronunciation Learning

Salah Eldeen Hamid<sup>1,3,4</sup> , Ossama Abdel-Hamid<sup>2</sup> , Mohsen Rashwan<sup>3,4</sup>

<sup>1</sup>Department of Electrical Engineering, Higher Technological Institute  
(10<sup>th</sup> of Ramadan City, Egypt)

<sup>2</sup>Department of information technology, Faculty of Computers and Information, Cairo University.  
Giza, Egypt.

<sup>3</sup>Department of Electronics and Communication Engineering, Cairo University.  
Giza, Egypt.

<sup>4</sup>The Engineering Company for the Development of Computer Systems; RDI.  
171 Al-Haram main st., 6<sup>th</sup> floor, Giza, Egypt  
{salah,mrashwan}@rdi-eg.com, ossama\_a@acm.org

## Abstract

This paper describes the use of a new automatic evaluation technique to evaluate speech-enabled computer aided pronunciation learning (CAPL) systems. The CAPL system is a part of a computer aided automatic training of correct recitation for the holy Qur'an for Arabic speakers.

The proposed technique evaluates the system by measuring the degree of usefulness of its feedback to learners. Evaluating CAPL systems by this means forces system designers to try to emphasize the system response 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 tunable thresholds and parameters.

## 1. Introduction

Due to the enormous achievements in computer technology and the general trend towards using computer based systems in educational systems, CAPL has received a considerable attention in recent years. Many research efforts have been done for improving such systems especially in the field of second language teaching (Herron et al 1999, Franco et al 1999; Cucchiarini et al 1998; Witt 1999). In this system we target automatic training for the correct recitation of the holy Qur'an for Arabic speakers. The shortage of experienced teachers in most of environments and/or lack of sufficient time at learner's side makes this system a highly demanded one. It not only helps students to learn how to recite the holy Qur'an but also helps them to correct their mistakes in formal Arabic pronunciation.

### 1.1 System Description

Figure (1) shows a block diagram of the on-line pronunciation error detection system. It deploys HMM-based speech recognizer that segments the utterance under test according to reference HMM acoustic models, available current learner adaptation data, and current verse pronunciation variants. The speech recognizer associates each decision it makes with a corresponding confidence score that is used to choose suitable feedback response to the learner.

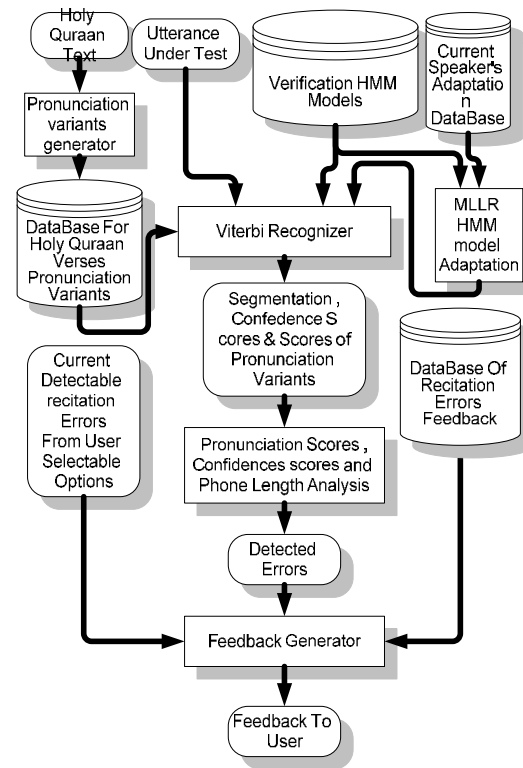


Figure (1) Block diagram of the on-line pronunciation errors detection system

When the system suspects the presence of a pronunciation error with low confidence score the system has some 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).

## 1.2 Need for Automatic Evaluation

This system needs a suitable evaluation technique which is used to check the effects of modification in any part of the system to achieve fine tuning for system parameters and as a measure of the system quality. As parameter tuning is a continuous and critical procedure to overall system performance; automating the evaluation process is crucial for system development.

## 1.3 Disadvantages of the Traditional Evaluation Techniques

In order to investigate some previously used evaluation techniques we used the traditional technique that is based on comparison of the human experts' transcription with speech recognizer selected pronunciation variant. This results in measuring the system accuracy in correct phoneme or word percentages.

Human experts sometimes disagree on one judgment on a phoneme pronunciation which can also happen for the same expert in different sessions. The major cause of disagreement is that 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 disregard an adjacent minor mistake.

Though we found this disagreement is less than 3% of the evaluation database, when the system approaches high accuracy decisions -in some specific problems-, 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.

Some other techniques are based on computing the correlation between human and computer ratings (Neumeyer et al 1996). In this technique the system feedback is just a rating of the quality of pronunciation of the utterance so it is not suitable for our target system.

In this paper we present a new evaluation technique, which tries to overcome these problems by measuring degree of usefulness of the system response to the user.

## 2. Evaluation database

The evaluation database contains a set of voice utterances representing the recitations of randomly selected users of the system of different gender, age and proficiency combinations.

These utterances was evaluated by a number of language experts, and labeled with the actual pronounced phonemes. Each expert was allowed to transcript 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.

This database is used to evaluate the system, by comparing the system responses with human experts' transcriptions.

So the database consists of a set of utterances, and all acceptable transcriptions for each utterance. It means that some pronounced segments may have more than one acceptable judgment.

The database is split into two parts

- a- Calibration set: for calculating optimum values for system parameters and calibrating confidence score thresholds.
- b- Evaluation set: used for the final evaluation of the system.

## 3. The Automatic Evaluation Procedure

As the system target is to be used by learners, so the system evaluation will be based on the degree of user benefit from the system responses. That's because the system deploys confidence score and there may be more than one accepted response. For example when the system suspects the presence of a pronunciation error with low confidence the system has many alternate responses as described in Sec. 1.1 of this paper.

### 3.1 Score matrix computation

For the evaluation database, the judgment has four possibilities:

- 1- Correct (accepted by all human experts).
- 2- Identified pronunciation error (all human experts reported the same error).
- 3- Not Perfect (human experts disagreed whether to reject or accept the pronunciation). That can happen when pronunciation is not perfectly correct.
- 4- Wrong with unidentified error type (human experts agreed that a pronunciation error exists but disagreed on its type). That can happen when the user makes complex or undocumented errors.

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 verse. 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.

So 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

For each possible system judgment, we define the system response score measuring usefulness of the response given the human experts' judgments to generate *score matrix* as shown in table (1).

Values of the used score matrix can vary with the progress level of the user, as advanced users can benefit from system responses that may annoy beginners and vice versa.

The score matrix is computed by presenting examples to a number of experts to set their acceptance rating for each system response for a number of utterances given a specified user level. They put their rating as a number from -10 to 10, where the highest rating is given for best responses (Correct for correct speech segments, or Wrong with the correct error type for wrong speech segments). These ratings will be averaged for each item in the score matrix as in table (1). Thus for each user level there can be a unique score matrix.

The confidence thresholds are set by searching for such thresholds that maximize the average system response score computed using the score matrix. It can be seen that each one of the three thresholds can be computed alone. For this purpose we use the data set prepared especially for the thresholds calibration.

The overall evaluation is done afterwards using the total response score computed using the evaluation data set.

## 4.Results

Table (2) shows a sample application of the evaluation system for novice users. The table shows the distribution of occurrences of judgment-response pairs in the matrix cells.

As we see in table (2), 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 52% of pronunciation errors, reported unidentified errors for 4% and gave "Repeat Request" for 24% of the errors. The system made false acceptance of 17% of total errors.

		Human judgement			
		Correct	Wrong	Not Perfect	Wrong with unidentified error types
System judgement	Correct	10	-10	10	-10
	Wrong with same error type	-10	10	10	10
	Wrong with wrong error type	-10	-3	-6	-3
	Repeat Request	-3	3	6	3
	Wrong with Undefined error	-6	6	3	6

Table 1 : Sytem response score matrix

		Human judgement				
System Judgement		Correct	Wrong	Not Perfect	Wrong with Unidentified Error Types	Total
	Correct	80.89%	1.43%	1.07%	0.00%	83.39%
	Wrong with same error type	0.00%	4.29%	0.18%	0.00%	4.46%
	Wrong With Wrong Error Type	0.00%	0.18%	0.00%	0.18%	0.36%
	Repeat Request	8.75%	1.96%	0.71%	0.00%	11.43%
	Wrong with Undefined Error	0.00%	0.36%	0.00%	0.00%	0.36%
	Total	89.64%	8.21%	1.96%	0.18%	100.00%

Table 2 : Occurrences Distribution

The average response score was 8.2. This high value is due to high correct acceptance percentage (80.9%). This average value is mainly used to compare two system settings in order to select the better setting with respect to response score.

### 5. Conclusion and Future Work

An algorithm for evaluating the response of CAPL systems that deploy confidence scores was implemented. The system features flexibility to combine multiple human experts' judgments to reach a subjective measurement of usefulness of the system response.

Another important advantage of the proposed algorithm, that it enables automation of the evaluation process. That enabled system developers to use it for tuning system parameters and in calculating confidence score thresholds.

CAPL systems must be evaluated at all possible inputs, which include incomplete utterances, hesitation, out of vocabulary words, very high noise levels and uncooperative users. That means the evaluation database should include utterances of those types. Also it means that human experts should have more evaluation options for speech segments. Also system responses should be increased to accommodate these situations to gain user trust.

### Acknowledgment

Special thanks are posed to The Engineering Company for the Development of Computer Systems (RDI) <http://www.rdi-eg.com> for its support of the pioneer application of CAPL technology in the holy Qur'an recitation learning. We gratefully acknowledge their support for this research.

We must mention valuable efforts of speech technology and linguistic support teams who helped build the evaluation database. Special thanks are posed to Waleed Nazeeh, Ahmad Ragheb, Naim Abdelghani and Badr Mahmoud.

### References

- Cucchiarini, C. & Strik, H. & Boves, L. (1998), Automatic pronunciation grading for Dutch. Proceedings STiLL '98, Marholmen, Sweden, pp.95-98 .
- Franco, H., Neumeyer, L., Ramos, M., and Bratt, H., (1999) Automatic Detection of Phone-Level Mispronunciation for Language Learning, Proc. of Eurospeech 99, Vol. 2, 851-854, Budapest, Hungary.
- Herron, D., Menzel, W., Atwell, E., Bisiani, R., Daneluzzi, F., Morton, R., Schmidt, J. A. (1999) Automatic localization and diagnosis of pronunciation errors for second-language learners of English, Proc. 6 th European Conference on Speech Communication and Technology, Eurospeech 99, Budapest, pp. 855-858.
- Neumeyer, L., Franco, H., Weintraub, M. and Price, P. (1996) Automatic text-independent pronunciation scoring of foreign language student speech. Proc. Proc. Int. Congress on Spoken Language Processing (ICSLP) '96,
- Witt, S. (1999) Use of Speech Recognition in Computer-Assisted Language Learning. PhD thesis, Cambridge University Engineering Department, Cambridge, UK.

This document was created with Win2PDF available at <http://www.daneprairie.com>.  
The unregistered version of Win2PDF is for evaluation or non-commercial use only.