Computer Aided Pronunciation Learning System Using Speech Recognition Techniques

Sherif Mahdy Abdou[24]*, Salah Eldeen Hamid*[34]*, Mohsen Rashwan*[14] *Abdurrahman Samir*[14]*,*
Ossama Abd-Elhamid[24]*, Mostafa Shahin*[14]*, Waleed Nazih*[24]

[1]Department of Electronics and Communication Engineering, Cairo University. Giza, Egypt.
[2]Department of IT, Faculty of Computers and Information, Cairo University. Giza, Egypt.
[3]Department of Engineering and Applied Sciences, Umm Al-Qura University, Mecca, Saudi Arabia
[4]Research & Development International (RDI®), Giza, Egypt
{sabdou, salah, mrashwan, asamir, ossama_a, mostafa_shahin, w_nazih}@rdi-eg.com

Abstract

This paper describes a speech-enabled Computer Aided Pronunciation Learning (CAPL) system HAFSS©. This system was developed for teaching Arabic pronunciations to non-native speakers. A challenging application of HAFSS© is teaching the correct recitation of the holy Qur'an. HAFSS© uses a state of the art speech recognizer to detect errors in user recitation. To increase accuracy of the speech recognizer, only probable pronunciation variants, that cover all common types of recitation errors, are examined by the speech decoder. A module for the automatic generation of pronunciation hypotheses is built as a component of the system. A phoneme duration classification algorithm is implemented to detect recitation errors related to phoneme durations. The decision reached by the recognizer is accompanied by a confidence score to reduce effect of misleading system feedbacks to unpredictable speech inputs. Performance evaluation using a data set that includes 6.6% wrong speech segments showed that the system correctly identified the error in 62.4% of pronunciation errors, reported "Repeat Request" for 22.4% of the errors and made false acceptance of 14.9% of total errors.

Index Terms: Pronunciation error detection, Qur'an recitation

1. Introduction

Computer Aided Pronunciation Learning (CAPL) has received a considerable attention in recent years. Many research efforts have been done for improvement of such systems especially in the field of second language teaching [1] [2].

A challenging application for CAPL is the automatic training for correct recitation of the holy Qur'an for Arabic speakers. In contrast to the foreign language training task, where a wide variety of pronunciations can be accepted by native speakers as being correct, the holy Qur'an has to be recited the same way as in the classical Arabic dialect and the tolerance for allowed variation is very fine.

There have been initial attempts to attack this problem. El-Kasasy [3] implemented a system that segments speech signal into syllabic units. Each test syllabic segment is compared with the reference one through dynamic time warping then the test syllabic segment is accepted or rejected as a whole. The system didn't give any detailed feedback about the type of error. Omar [4] proposed to use hidden Markov model based speech verification system. In that work the Arabic phoneme set was clustered to a group of clusters and the pronunciation assessment was accomplished in two steps: First, the input speech is segmented into a sequence of these clusters of phonemes. Substitution, insertion and deletion errors are detected in this stage by comparing the sequence of detected clusters with the given hypothesized sequence. Second, these units are tested by discriminatively trained HMM models. This hierarchal system added complexity and its performance was worse than purely statistical approaches like HMM-based systems. To the best of our knowledge, there isn't any reported implementation of a system for the automatic assessment of recitation of the Holy Qur'an. In this paper we introduce "HAFSS©", a commercial product resulted from several years of research and development efforts at RDI®. It is used for teaching Arabic pronunciations to non-native speakers. We will show how HAFSS© can be used for the challenging task of teaching the correct recitation of The Holy Qur'an. We will show how it can assess the quality of a user's recitation and produce a feedback messages to help him locate his mispronounced letters and eventually overcome them.

In the following sections of this paper, section 2 includes a description of the HAFSS© system architecture. Section 3 describes the pronunciation hypotheses generator module. Section 4 describes the confidence scoring module. Section 5 includes recent improvements in HAFSS© acoustic models and the system evaluations results. Section 6 includes conclusions.

2. System description

Figure 1.Shows the block diagram of the HAFSS© system. Its main blocks are:

1. **Verification HMM models**: Is the acoustic HMM models for the system.
2. **Speaker Adaptation:** Is used to adapt acoustic models to each user acoustic properties in order to boost system performance. It uses speaker classification, Maximum Likelihood Linear Regression (MLLR) speaker adaptation algorithms and supervised incremental technique [6].
3. **Pronunciation hypotheses generator**: It analyzes current prompt and generates all possible pronunciation variants that are fed to the speech recognizer in order to test them against the spoken utterance. It is described in details in section 3.

4. **Confidence Score Analysis:** It receives n-best decoded word sequence from the decoder, then analyzes their scores to determine whether to report that result or not. It is described in details in section 4.
5. **Phoneme duration analysis:** For phonemes that have variable duration according to its location in the Holy Qur'an, this layer determines whether these phonemes have correct lengths or not. To overcome inter-speaker and intera-speaker variability in recitation speed that may mislead the phone duration classification module. An algorithm for Recitation Rate Normalization (RRN) was developed [5].
6. **Feedback Generator:** Analyze results from the speech recognizer and user selectable options to produce useful feedback messages to the user.

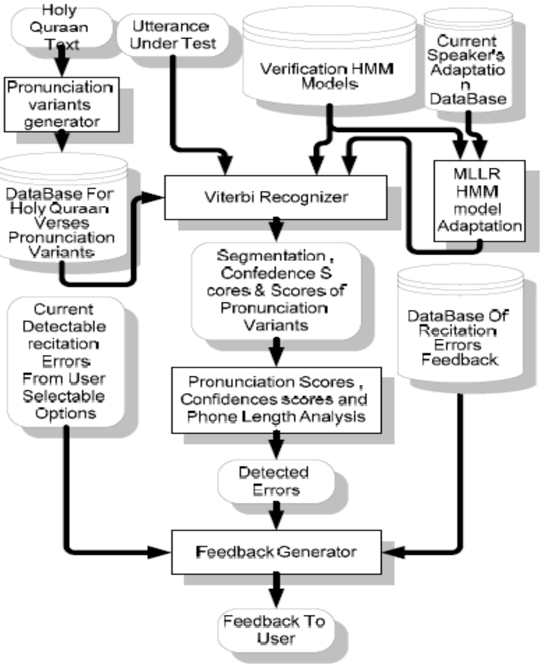

Holy Quraan Text
Utterance Under Test
Verification HMM Models
Current Speaker's Adaptatio n DataBase
Pronunciation variants generator
DataBase For Holy Quraan Verses Pronunciation Variants
Viterbi Recognizer
MLLR HMM model Adaptation
Segmentation , Confedence S cores & Scores of Pronunciation Variants
Current Detectable recitation Errors From User Selectable Options
DataBase Of Recitation Errors Feedback
Pronunciation Scores , Confidences scores and Phone Length Analysis
Detected Errors
Feedback Generator
Feedback To User

Figure 1 *Block diagram of HAFSS©*

3. Generation of pronunciation errors hypotheses

The pronunciation error hypotheses in HAFSS© are represented in the form of a linear lattice that is flexible enough to support error hypothesis addition, deletion and overlapping of probable mispronunciations. Figure 2. shows a block diagram for the search lattice generation module.

The lattice generator is built on basis of RDI® holy Qur'an transcription engine. The transcription engine is built in the form of multi-layer event driven modules. This architecture was selected to enable any higher level analysis module to use this core engine and benefit from its results. That what was actually done with our pronunciation variants lattice generator.

The events engine scans the input holy Qur'an Ottoman text searching for symbols and features and at each probably pronounced character it generates its code, its pronunciation status and its acoustic characteristics (such as voicing, place of articulation, nasalization and aspiration). The transcription engine analyzes those codes and characteristics and generates the corresponding correct phonetic transcription according to the holy Qur`an recitation rules and their exceptions. The pattern engine gathers all the information from proceeding layers and generates pronunciation patterns at probable pronunciation locations. These pronunciation patterns are used for matching with pronunciation variants rules at the lattice generator. The lattice generator sorts the matched rules descending with their error relevance or impact then all rules resulting in the same phoneme sequence are omitted except the first one. Finally the lattice is generated with remaining pronunciation variants in a format suitable to the speech recognizer. Also a mapping file is generated that holds the locations of the suitable feedbacks for each pronunciation variant.

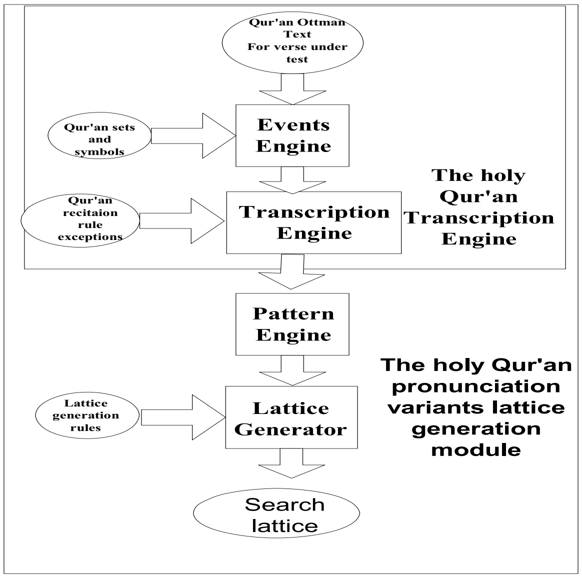

Qur'an Ottman Text For verse under test
Qur'an sets and symbols
Events Engine
Qur'an recitaion rule exceptions
Transcription Engine
The holy Qur'an Transcription Engine
Pattern Engine
Lattice generation rules
Lattice Generator
The holy Qur'an pronunciation variants lattice generation module
Search lattice

Figure 2. *Block diagram for the holy Qur`an pronunciation variants lattice generator*

Because we needed our system to generate user helpful feedbacks and that most Qur`an learners are not familiar with phonemes we choose our lattice unit similar to the one used in traditional methods of the holy Qur`an recitation teaching. In those methods the text is divided to basic units Consonant + short vowel (CV), Consonant + long vowel (CVV), Non-vowelled consonant (C), Repeated consonant +short vowel (CCV), Repeated consonant + long vowel (CCVV), Non-vowelled repeated consonant (CC).

Studying the types of recitation errors we found that they usually depend mainly on current unit, the nearest proceeding unit and the nearest succeeding unit. Effect of far units is limited to the error rank and can be ignored safely on the assumption that all pronunciation hypotheses are generated. Figure 3. shows the rules template that we used.

We managed to construct a database of 663 rules of pronunciation errors in the holy Qur`an recitation. This is the database of recitation errors that we have recognized; the architecture of the system enables us to freely use a subset of this database to produce limited scale realizations of the system. A probable application is to generate "Vowel Type" errors only for a realization dedicated to beginners. This database is connected via error code to two other databases. The first database is the feedback database which holds the coloring codes, the readable feedbacks and the audible feedbacks. The second database also holds links between each specific recitation error and relevant holy Qur`an recitation rules. This database is used to filter the pronunciation errors to concentrate on specific recitation rules for a given lesson.

```
IF For Current unit
    Phoneme = $Ph_c$ , Vowel type = $V_c$ , Vowel length =$L_c$ ,
    Doubled = $S_c$, Pronounced = $P_c$ and Concealment =$E_c$
AND For previous unit
    Phoneme = $Ph_p$ , Vowel type = $V_p$ , Vowel length =$L_p$ ,
    Doubled = $S_p$, Pronounced = $P_p$
AND For next unit
    Phoneme = $Ph_n$ , Vowel type = $V_n$ , Vowel length =$L_n$ ,
    Doubled = $S_n$, Pronounced = $P_n$
THEN Add path to recitation error lattice with  parameters
    Error Code = $C$, Error Type = $T$, Error word =$W$,
    Error frequency rate = $F$, Phoneme = $Ph$,
    Vowel type=$V$, Vowel length = $L$, Doubled = $S$,
     Pronounced = $P$ and Concealment = $E$
Where $S$, $P$ and $E$ are binary flags.
```

Figure 3. *The rules templat*e

4. Confidence scoring

The speech recognizer in HAFSS© associates each decision it makes 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 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).

The implemented confidence scoring in the system is based on the Likelihood ratios [5] 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, $N = E – S$.

Due to the fact that 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 center of gravity of the two probability distribution.

5. System evaluations

To reach an evaluation for the complete HAFSS© system, an evaluation database was collected. This test set consisted of 507 utterances representing the recitations of randomly selected users of the system of different gender, age and proficiency combinations. These utterances were evaluated by language experts and labeled with the actual pronounced phonemes then used to evaluate the system, by comparing the system responses with human experts' transcriptions.

This evaluation procedure was used to adjust system parameters by studying the change of overall performance due to changes of each parameter. The HMM used in HAFSS© is triphone tied state model. Each state is modeled by mixture of Gaussians. We run several experiments to tune the parameters of the acoustic model which are the clustering threshold th_c, the outlier threshold th_o, and the number of mixtures N_{mix}. Table .1 shows the results of these tuning experiments.

Table 1. *Model tuning results*

N_{mix}	Th_c	Th_o	*Num. Clusters*	*Num. Gaussians*	*C.J.*
4	350	100	3829	15316	96.96 %
4	350	150	3597	14388	97.19 %
6	350	150	3597	21582	97.34 %
6	500	150	2899	17394	97.38 %
8	500	150	2899	23192	97.21 %
8	800	150	2152	17216	97.58 %
SAT			2152	17216	98.22 %

Where *CJ* is the Correct Judgment ratio of HAFSS©, i.e accept correctly pronounced phones or report same pronunciation error as the human expert. From results in table.1 we can see that the tuned model gave 0.63% relative improvement in the system judgments compared to the initial model. This tuned model was further improved using Speaker Adaptive Training (SAT) [6] which is a method that integrates adaptation in the training phase. To build the SAT model we used a set of 38 speakers, of different edge and gender, and included 11600 utterances. Results in Table .1 show that with the SAT model we got another 0.62% relative improvement in system judgments.

In the second experiment we deployed the confidence score in generating the system response. The system could be any one of the four alternatives described in section 4. Table 2. shows the results after using confidence score. Results in tables 1 and 2 are phone-based measures. A more accurate measure for the system performance should be based on the degree of user benefit from the system responses. To calculate this measure we considered the system response messages for the utterance as a whole rather than for each phone separately. Table .3 shows the message-based system evaluation results.

As we see in table .3, for correct speech segments the system yielded "Repeat Request", which asks the user to try again, for about 5.1% of the total correct utterances. 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 6.6% of the data, the system correctly identified the error in 62.4% of pronunciation errors and reported "Repeat Request" for 22.4% of the errors. The system made false acceptance of 14.9% of total errors.

Table 2. System performance with confidence.

		Human judgment	
System judgment		Correct	Wrong
	Correct	96.87	0.04
	Wrong with same error type	1.16	0.59
	Wrong with wrong error type		0.04
	Repeat Request	0.46	0.02
	Unidentified error	0.74	0.06

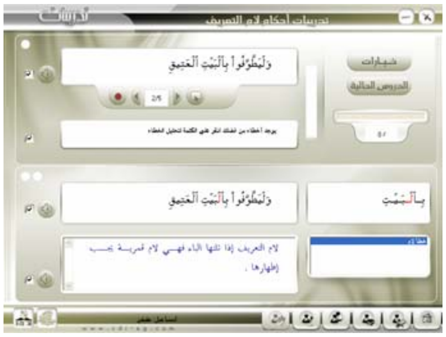
Table 3. *Message based system performance with confidence*.

		Human judgment	
System judgment		Correct	Wrong
	Correct	80.13	0.99
	Wrong with same error type	6.29	4.14
	Wrong with wrong error type		0
	Repeat Request	4.80	1.49
	Unidentified error	2.15	0

6. Conclusions

In this paper we introduced the HAFSS© system. This system is of multimedia type with pre-recorded Holy Qur'an recitations, recitation teaching text materials and teaching animations. Figure 4 shows two screen captions from HAFSS© system for the exercises screen and the pronunciation learning flash.

The HAFSS© system proved to be a useful one for the challenging task of automatic training for the correct recitation of the holy Qur'an for Arabic speakers. 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. The shortage of experienced teachers in most of environments and/or lack of sufficient time at learners' side makes this system a highly demanded one. Our future work on HAFSS© will focus on using discriminative training techniques to improve the discrimination between some confusable pronunciation alternatives. Also we will try to improve the confidence score by using some articulation features.



وَلْيَطَّوَّفُوا بِالْبَيْتِ الْعَتِيقِ

يوجد أخطاء من فضلك انقر على الكلمة لتحليل الخطاء

وَلْيَطَّوَّفُوا بِالْبَيْتِ الْعَتِيقِ

بِالْبَيْتِ

لام التعريف إذا تلتها الباء فهي لام قمرية يجب إظهارها .

إظهار

Figure .4.a: *Exercise screen*

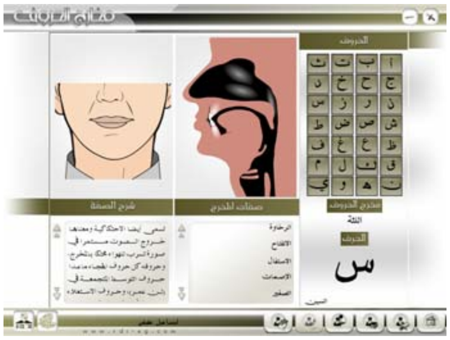


Figure .4.b: *Pronunciation learning flash*

7. Acknowledgments

Special thanks are posed to The Research & Development International Company (RDI®) for its support of the pioneer application of CAPL technology in holy Qur`an recitation learning. We must also mention valuable efforts of speech technology and linguistic support teams at RDI®.

8. References

[1] 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.

[2] Witt, S. (1999) Use of Speech Recognition in Computer-Assisted Language Learning. PhD thesis, Cambridge University Engineering Department, Cambridge, UK

[3] El-Kasasy, M. S., "An Automatic Speech Verification System", Ph.D. Thesis, Cairo University, Faculty of Engineering, Department of Electronics and Communications, Egypt, 1992.

[4] Omar, M. K., "Phonetic segmentation of Arabic speech for verification using HMM", M.Sc. Thesis, Cairo University, Faculty of engineering, Department of Electronics and Communications, Egypt, Jan. 1999.

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

[6] C. J. Leggetter, "Improved Acoustic Modeling for HMMs using Linear Transformations," Ph. D. Thesis, University of Cambridge, Cambridge, UK, 1996.