Enhancing usability of CAPL system for Qur'an recitation learning

Abdurrahman Samir [1*], *Sherif Mahdy Abdou* [23], *Ahmed Husien Khalil* [4], *Mohsen Rashwan* [24]

[1] Katholieke Universiteit Leuven - Dept. ESAT, Belgium
[2] Research & Development International (RDI®), Giza, Egypt
[3] Department of IT, Faculty of Computers and Information, Cairo University. Giza, Egypt
[4] Department of Electronics and Communication Engineering, Cairo University. Giza, Egypt
asamir@esat.kuleuven.be, {sabdou, ahmed.husien, mrashwan} @rdi-eg.com

Abstract

This paper describes some enhancements for a speech-enabled Computer Aided Pronunciation Learning (CAPL) system HAFSS©. This system was developed for teaching Holy Qur'an recitation rules and Arabic pronunciations to non-native speakers. One important point that is critical in any practical language learning system that exploits ASR technology is the user enrolment time. In this paper we introduce the modifications that were done on the baseline system to reduce the amount of the enrolment time while keeping the system accuracy at the same level. Also we introduce results of some experiments that measure the correlation between the judgments of HAFSS system and the judgments of human experts. Also we measured the usefulness of HAFSS system for beginner users by measuring their proficiencies before and after using the system.

Index Terms: pronunciation learning, Arabic Language, usability, Speaker Adaptive Training.

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], [3], [4]. A challenging application for CAPL is the automatic training for correct recitation of the holy Qur'an for Arabic speakers [5], [6]. 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 [7], [8]. In [5] a commercial system for automatic assessment of recitation of the Holy Qur'an "HAFSS©" developed at RDI was presented. Working for a commercial system drives us to give much effort to usability issues while tackling the CAPL problem. One important point that is critical in any practical language learning system that exploits ASR technology is the user enrolment time. Many applications, as in [1], prefer to exclude this phase entirely working with speaker independent model that is tuned to give good performance without adaptation. To reduce user enrolment time in HAFSS© system we made several modifications. The first one is utilizing Speaker Adaptive Training (SAT) [9] which allowed for achieving nearly the same level of accuracy with very small amount of adaptation data compared to the baseline system. Although the user enrolment time was reduced significantly, some users wanted to do it in several sessions rather than being forced to complete it in one session. Cascade adaptation was proposed to allow for multi-session adaptation meanwhile it produces almost the same transformation matrix of the single session adaptation. The baseline system restricted adaptation to utilize only the correctly pronounced utterances, the ones that matched the reference prompts. This resulted in very few selected sentences in the short-enrolment scenario. We proposed using a confidence score to increase the number of utterances utilized for adaptation.

Evaluation of the overall system performance is done by measuring the recognition accuracy and segmentation accuracy of the system. Also a novel Phone segmentation accuracy measure was proposed. As pronunciation judging does not have a clear-cut, some pronunciations seen as errors by one judge can be seen as correct by another. We measured the correlation between HAFSS© and four experts judgments on one hand, and between judgments of the four experts themselves on the other hand. Also we measured the usefulness of HAFSS system for beginner users by measuring their proficiencies before and after using the system.

In the following sections of this paper, section 2 includes a description of the baseline system. Section 3 discusses changes made in the user enrolment scenario. Section 4 describes the cascade adaptation idea. Section 5 describes the segmentation accuracy measure. Section 6 includes descriptions of evaluation databases and the experimental results. Section 7 includes conclusions.

2. Baseline system description

Figure 1 Shows the block diagram of the HAFSS© system [5], [6]. 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 [9].

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.

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 as described in [5].

5. Phoneme duration analysis: For phonemes that have variable duration according to its location in the Holy Qur'an,

* This research was fully developed in RDI® (http://www.rdi-eg.com) speech labs.

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 [6].

6. Feedback Generator: Analyze results from the speech recognizer and the user selectable options to produce useful feedback messages to the user.

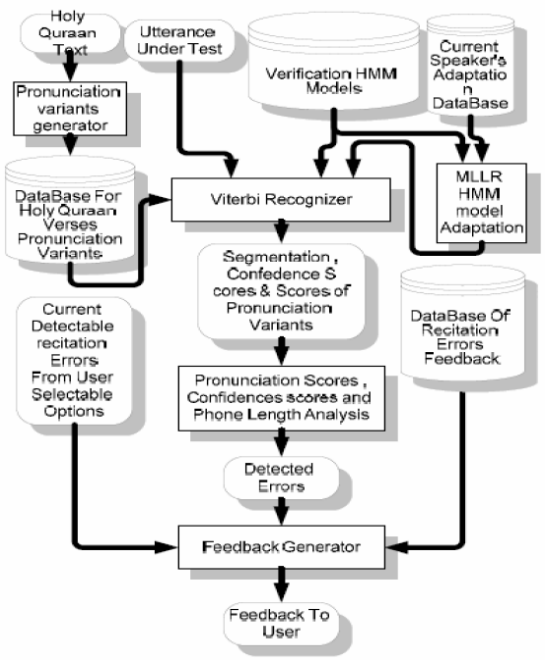

Figure 1: the baseline system block diagram

3. Changes in user enrolment scenario

Reducing enrolment time of the CAPL systems is of great importance [1]. To achieve this goal, some changes in user enrolment scenario have taken place:

1. Enrolment data selection: In the original system, the user was asked to utter carefully selected words and sentences from the Holy Qur'an. But those selected sentences don't come in sequence in the Holy Qur'an reading, which make users lose the expectation of the next sentence to be uttered (this expectation was shown to improve users performance). So a new enrolment prompt data was prepared that take data in sequence from the same part of the Holy Qur'an and also they are easy to utter for beginner reader.

2. In the baseline enrolment process, user was prompted to say the prompt sentence again if the speech verification engine detects errors in his recitation. This attitude has proven to be inefficient from two viewpoints: First, from the educational viewpoint, this gives the new system user wrong feeling about the system performance especially if he pronounced the utterance correctly while the system judged it as error as a result of the weakness of the initial model used for verification until the user personal profile is built. Even if the user wrongfully pronounce the utterance, it is not advisable to tell him to reenter the sentence again as he hasn't started his learning process yet. Secondly, refused utterances most probably will be refused again as the initial user model isn't mature enough to give accurate decisions concerning user recitation.

3. Due to the fact that the user initial model isn't accurate enough to catch user pronunciation errors, the baseline design of the user enrolment depends on decoding the same utterance with more than one model. This is better from the system accuracy prospective, but it isn't good from the usability point of view as the system loses the advantage of decoding the utterance live while it is being said. So, each utterance would be decoded live once and then decoded one or two times again which consumes long time in the enrolment process. It is better to relay on single initial model for the user while trying to solve the accuracy issue by other means (e.g. involvement of confidence in the enrolment phase).

4. Correctness of utterances was the only criteria used for selecting adaptation data. It is proposed to involve confidence scoring in the enrolment process. 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 recitation phone sequence, if we have high confidence in the decoder output.

4. Cascade Adaptation

As users exercise recitation rules using HAFSS©, users' profiles are being enhanced using collected data for each user because speakers' transformations are getting more accurate. The baseline system [5] uses all data collected from certain user to create the new transformation. Although this improves system performance gradually, it is a time consuming process. Alternatively, cascade adaptation is proposed to improve the processing time with minimal degradation in performance. The new transformation will be calculated 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 decoding time (as the number of transformations will be incremented for each adaptation). This problem is solved using transformation summation as shown in the following equations. If we have two transformations T1 and T2, where T2 is calculated based on T1. Substituting from (1) into (2) to get the relation between new and old models in terms of components of transformation A and b. Equations (4) and (5) are used to calculate the summation transformation components.

$$\mu_1 = A_1\mu_o + \mathrm{b}_1 \quad (1)$$

$$\mu_n = A_2\mu_1 + \mathrm{b}_2 \quad (2)$$

$$\mu_n = A_2\,(A_1\mu_o + \mathrm{b}_1) + \mathrm{b}_2 \quad (3)$$

So,

$$A_{sum} = A_2A_1 \quad (4)$$

$$b_{sum} = A_2\mathrm{b}_1 + \mathrm{b}_2 \quad (5)$$

As new and old transforms are built using different amounts of data, so it is expected that regression trees – used to compute these transforms – grow to different levels. A recursive tree search was build to find the association between transforms.

Transformation summation is tested by measuring the Euclidean Distance between two models, one is transformed by T1 and T2 respectively and the other is transformed by the augmented transform. The models differences were within 10^{-4} % of the expected model values. It worth to note that limited data amounts used to calculate each transformation might prevent the regression tree to grow so as to capture fine details of the speaker. So a threshold is put on the minimum amount of data used to get the transformation.

5. Segmentation Accuracy measure

As the system process the testing corpus, it produces an automatic segmentation based on the input exercise lattice and the acoustic model. Figure 2 shows the HAFSS© system segmentation versus the human expert segmentation. Segmentation accuracy is measured using the following formula:

Seg_Acc= ((Overlap/Human_Len)+(Overlap/Hafss_Len))/2

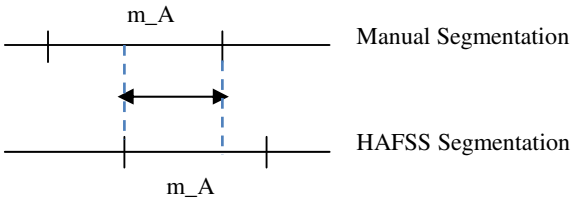

Figure 2: Hafss segmentation versus manual segmentation

For a phone to get the highest accuracy, phone boundaries, in system segmentation, must be very close to boundaries of manual segmentation. So this measure could handle cases where the system segmentation lays inside the manual segmentation and vice versa.

6. System Evaluation and Experiments

Evaluation databases contain utterances representing the recitations of randomly selected users of the system of different gender, age and proficiency combinations. These utterances were evaluated by several language experts, and labeled with the actual pronounced phonemes. For ambiguous speech segments experts were allowed to mark them as not clear segments. The availability of standard database is essential to compare different systems [10]. In this work two testing corpuses were used:

1. RDI-2004TDB Testing Corpus: It is used to evaluate system segmentation accuracy as it was manually segmented and transcribed. It consists of 512 utterances from 12 speakers (5 males, 2 females, 5 children). The total number of phones in this corpus is about 5,000 phones. The small size of this corpus is an obvious disadvantage. Manual segmentation is a costly process to be done on large database.

2. RDI-2006aTDB Testing Corpus: It is used for all experiments done on the system (except the system segmentation accuracy). It is transcribed then revised by Holy Qur'an recitation experts. It consists of 2230 utterances from 31 speakers (17 males, 8 females, 6 children). The total number of phones in this corpus is about 27,000 phones.

The proposed evaluation is composed of two main parts:
1. Measuring the system phone segmentation accuracy. This gives insight on the acoustic model strength. If the system segmentation accuracy was found to be lower than certain accuracy level, system couldn't judge accurately the correctness of utterances, as the segmentation is a basic block for all succeeding steps of pronunciation assessment.

2. Measuring the system capability to discriminate between different phones.. System classification accuracy is measured on RDI-2006aTDB Holy Qur'an recitation database. In our case, there won't be deletion or insertion errors (as in speech recognition evaluations) because the input lattice force the decoder to have limited set of alternatives for each speech segment.

Effect of using SAT on reduction of user enrollment time was evaluated by measuring system recognition accuracy with various adaptation data sizes both with and without SAT.

Adapt. data size	50 sec.	100 sec.	450 sec.
Without SAT	86.30	87.856	90.05
With SAT	89.81	91.29	91.98

Table 1: System recognition accuracy with variable data sizes

Without using SAT, we should collect 450 seconds in the adaptation phase to achieve 90 % starting accuracy. After using SAT 100 seconds are being collected to achieve the same accuracy level as shown in table 1.

Confusion matrices both with and without SAT are shown in Table 4, 5 with 100 seconds of adaptation data for all speakers. For basic training (before SAT), the core model was built using single speaker data. The reference speaker data size is 3 hours with 14% of the data size was manually segmented to initialize the acoustic model parameters. For SAT training, the previously mentioned SD model was used as the seed model. 80 Speakers were used for SAT model building (about 45 hours of audio data).

		Human judgment		
		Correct	Wrong	Not Clear
HAFSS	Correct	87.78	0.61	1.7
	Wrong with correct error	7.91	1.17	0.66
	Wrong with wrong error		0.17	

Table 2: Phone recognition accuracy without SAT

		Human judgment		
		Correct	Wrong	Not Clear
HAFSS	Correct	90.16	0.68	1.84
	Wrong with correct error	5.53	1.15	0.53
	Wrong with wrong error		0.14	

Table 3: Phone recognition accuracy with SAT

After using SAT, average phone segmentation accuracy was enhanced for more than 75% of phones. Segmentation accuracy of some phones such as Zaa' /~z/, Daad /~d/ was reduced by 20-30%, and for the other phones it was reduced by less than 5%. Best segmentation accuracies were achieved by long vowels (either before of after applying SAT).

In the second experiment we measured the correlation between HAFSS© and four experts judgments on one hand, and between judgments of the four experts themselves on the other hand. 300 utterances was collected from a non-proficient user, then it was judged on utterance level by HAFSS© and the four experts.

	HFS	Expt1	Expt2	Expt3	Expt4
HFS		0.78	0.77	0.89	0.82
Expt1	0.78		0.82	0.78	0.77
Expt2	0.77	0.82		0.75	0.75
Expt3	0.89	0.78	0.75		0.82
Expt4	0.82	0.77	0.75	0.82	

Table 4: Correlation between HAFSS and human experts

Table 4 shows that HAFSS© average correlation with other experts is 0.815 while the average correlation between experts themselves (excluding HAFSS) is 0.78. This result shows that the accuracy of HAFSS system can be considered in the range of human perception of pronunciation errors.

In the last experiment we measured the usefulness of the HAFSS system to beginner users. Three rules were selected that are known to be problematic to new learners. 10 users were encouraged to practice these three rules using HAFSS©. Learner's proficiency was tested before and after using HAFSS©. The learners progress was significant in the first two hours (focusing on a single rule) while it becomes small afterwards. Table 5 shows average progress for each tested rule.

	Before using HAFSS	After one hour	After two hours
Al Kalkala (vibration)	40 %	60 %	72 %
Edgham (Assimilation)	52 %	59 %	74 %
Al Eklap (Turning)	50 %	72 %	79 %

Table 5: Effect of using HAFSS on beginner users

7. Conclusions

In this paper we introduced some modifications to an existing CAPL system HAFSS©. The main target of such modifications was to enhance system usability by reducing the enrollment time. Using Speaker Adaptive Training (SAT) significantly reduced the needed adaptation data size from 450 sec. to 100 sec. with keeping the same level of system accuracy. Multi-session enrolment scenario was introduced utilizing cascade adaptation to give more flexibility for the system users without affecting the computed models. Involvement of a confidence scores while selecting adaptation data allowed the increase of the adaptation utterances. The HAFSS© system performance was tested against human experts. It was found that system performance falls in the range of human perception of pronunciation errors. Finally, the degree of usefulness of using HAFSS© for beginner users was tested. The learning curve increases fast in the first two hours of learning certain recitation rules then it becomes smaller. In our future work, we intend to use "user clustering" to boost the initial user model. Also, Techniques will be developed to overcome segmentation inefficiency resulted from using SAT.

8. Acknowledgements

Special thanks are posed to The Research & Development International Company (RDI®) for its support of this work. Authors appreciate Dr. Salah Eldeen Hamid keen efforts to build the baseline system. Also we would like to thank Mostafa Shahin and all speech technology and linguistic support teams at RDI® for their valuable efforts. We awe many thanks to Mr. Ahmad Ragheb, Mr. Yasser Fayez, and Dr. Saad Ryad for providing us with the usability experiments results.

9. References

[1] H. Franco et al., "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning", Proc. of InSTIL, Scotland, 123-128, 2000.

[2] S.M. Witt, S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning", Speech Communication 30, 95-108, 2000.

[3] Menzel, W., Herron, D., Bonaventura, P., and Morton, R. (2000). "Automatic detection and correction of non-native English pronunciations", Proceedings of InSTILL 2000, Dundee, Scotland, 49-56.

[4] Mak, B., Siu, M.H., Ng, M., Tam, Y.C., Chan, Y.C., Chan, K.W., Leung, K.Y., Ho, S., Chong, F.H., Wong, J., Lo, J. (2003). "PLASER: Pronunciation Learning via Automatic Speech Recognition", Proceedings of HLT-NAACL 2003, Edmonton, Canada, 23-29.

[5] S. Abdou, S. Hamid, M. Rashwan, A. Samir, O. Abdel-Hamid, M. Shahin, W. Nazih, "Computer Aided Pronunciation Learning System Using Speech Recognition Techniques", INTERSPEECH 2006 - ICSLP, Pittsburgh, PA, USA.

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

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

[8] 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.

[9] Gales, M. J. F. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report CUED/F-INFENG/ TR291, Cambridge University.

[10] K. Lee, A. Hagen, N. Romanyshyn, S. Martin, and B. Pellom, "Analysis and detection of reading miscues for interactive literacy tutors," in Proc. International Conference on Computational Linguistics, Coling, Geneva, Switzerland, August 2004.