Diagnostic CALL tool for Arabic learners

Majed Alsabaan[1] and Allan Ramsay[2]

Abstract. Our proposed work is aimed at teaching non-native Arabic speakers how to improve their pronunciation. This paper reports on a diagnostic tool for helping non-native speakers of Arabic improve their pronunciation, particularly of words involving sounds that are not distinguished in their native languages. The tool involves the implementation of several substantial pieces of software. The first task is to ensure the system we are building can distinguish between the more challenging sounds when they are produced by a native speaker, since without that, it will not be possible to classify learners' attempts at these sounds. To this end, we carried out a number of experiments with the well-known speech recognition Hidden Markov Model Toolkit (HTK), in order to ensure that it can distinguish between confusable sounds, such as the ones that people have difficulty with. Our diagnostic tool provides feedback in three different forms: as an animation of the vocal tract, as a synthesised version of the target utterance, and as a set of written instructions. We have evaluated the tool by placing it in a classroom setting, asking 40 Arabic students to use the different versions of the tool. Each student had a thirty minute session with the tool, working their way through a set of pronunciation exercises at their own pace. Preliminary results from this pilot group show that their pronunciation does improve over the course of the session.

Keywords: language learning, pronunciation support, articulation, non-native speaker, Arabic, speech recogniser, animated head, synthesised speech.

[1] malsabaan@cs.man.ac.uk.

[2] allan.ramsay@cs.man.ac.uk.

How to cite this article: Alsabaan, M., & Ramsay, A. (2014). Diagnostic CALL tool for Arabic learners. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds), *CALL Design: Principles and Practice*; *Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands* (pp. 6-11). Dublin: Research-publishing.net. doi:10.14705/rpnet.2014.000186

1. Introduction

Nowadays, learning a foreign language is an essential activity for many people. Proficiency in a foreign language is based on four different skills; reading, listening, writing, and pronunciation or speaking skills. We are particularly interested in pronunciation skills, and this paper aims to describe an attempt to give non-native Arabic speakers pronunciation support on how to make the different sounds that make up Arabic (i.e. help them sound more like native speakers) by giving them three forms of feedback: (1) an animation of the vocal tract corresponding to both the sounds learners have made and the sounds they should have made (i.e. the correct sounds), (2) a synthesised version of both what they said and what they should have said, and (3) an explanatory text of how they can pronounce the target sample correctly.

This project uses a speech recogniser called the HTK (Young et al, 2006) to identify the properties of speech signals for both native and non-native speakers. We trained the HTK to recognise phonemes of the input speech in order to obtain a phonetic analysis which has been used to give feedback to the learner. The HTK analyses the differences between the user's pronunciation and that of a native speaker by using *minimal pairs*, where each utterance is treated as coming from a family of similar words (i.e. two words with different meanings when only one sound is changed). This enables us to categorise learners' errors; for example, if someone is trying to say *cat* and the recogniser determines they have said cad, then it is likely that they are voicing the final consonant when it should be unvoiced.

Extensive testing shows that the system can reliably distinguish such minimal pairs when they are produced by a native speaker, and that this approach does provide effective diagnostic information about errors. In this way, we can provide feedback on acoustic data which we hope will enable learners to adjust their pronunciation.

The novel aspect about the current work is that (1) we will be applying these notions to Arabic, for which speech recognition is inherently harder than for English, (2) we will be giving multiple kinds of feedback to the learners and allowing them to choose between them, and (3) since we are providing multiple forms of feedback, we can evaluate the comparative effectiveness of each of them.

The evaluation has been done from three points of view: quantitative analysis, qualitative analysis, and questionnaire. Firstly, the quantitative analysis provides raw numbers indicating whether a learner is improving his/her pronunciation or not. Secondly, the qualitative analysis shows a behaviour pattern of what a learner did and how he/she used the tool. Thirdly, the questionnaire gives us a feedback from a learner and his/her comments about the tool.

2. Method

2.1. HTK experiments

The HTK experiments are a major stage to help determine what will be needed in order to animate the vocal tract and synthesise the phoneme sequence as feedback to learners. In our experiments, we used isolated words for the training samples. These samples were chosen by using the minimal pair technique. This technique helped learners distinguish between similar and problematic sounds in the target language (i.e. Arabic) through listening discrimination and spoken practice.

We investigated many factors (gender, words as terminal symbols, phonemes as terminal symbols) in order to improve the recognition accuracy as we reached an accuracy of 77%. Using these experiments, the HTK, with this acceptable recognition accuracy, was embedded into our Computer Assisted Language Learning (CALL) tool, giving us what a learner said while he/she was using our CALL tool.

2.2. The morphing

Prior to morphing, we had to draw the vocal tract for each phoneme. This drawing was done by tracing a set of pictures of Arabic phonemes and getting their coordinates. After that, we morphed from one drawing to another, in which we assigned phonetic units to individual snapshots. In more detail, we generated a set of morphs from a set of drawings using a hash table written in Java. This table, containing all the geometries that made up the animation we were trying to obtain, was generated using Dynamic Time Warping (DTW), which matched points in the two images to be morphed. This alignment must be done because the morphing cannot be carried out unless all morphed pictures have the same length sequences (i.e. same number of points). The animation of these pictures was done using Java3D.

Other researchers have done similar work using an animated head for this kind of task, but the underlying technology that they use is different (Liu, Massaro, Chen, Chan, & Perfetti, 2007; Massaro, 2004; Massaro & Cohen, 1995; Massaro & Light, 2003, 2004; Ouni, Cohen, & Massaro, 2005).

2.3. Speech synthesis

We used synthetic speech as one source of feedback to learners. To generate synthetic speech from a set of phonetic transcriptions, we used a tool called MBROLA. This tool is a text-to-speech system for most languages including Arabic. We supplied MBROLA with the phonemes of each word used in the training samples. MBROLA converted the script of word phonemes into speech. However, this speech sounds robotic and flat because we have a fixed length for each phoneme and a fixed pitch. To accurately estimate the length for a phoneme, we used both the alignment with real speech and a tool called Praat in order to get the phone length right. To improve pitch variations, we used a stress assignment program (Ramsay, Alsharhan, & Ahmed, 2014) which helped to find the stress of a given word and add the right pitch values. By having appropriate values of both phone length and pitch, the generated speech from MBROLA became more realistic, helping learners to listen to the correct pronunciation clearly.

2.4. The integration

The integration step is the last step for getting our diagnostic tool. In this step, we carried out a major piece of implementation which was done by integrating the two existing pieces of software. The first piece of software was for identifying mispronunciation by driving the speech recogniser (i.e. the HTK) in which we used phonemes as terminal symbols instead of words. The second piece of software was for animating a sequence of images of articulatory positions and performing a speech synthesis.

3. Discussion

In general, CALL tools are very difficult to evaluate because learning a language is a long, slow process and measuring the effect of a tool requires carrying out a large scale longitudinal study. The designers of Baldi (Massaro & Cohen, 1995) suggest that people’s pronunciation does, in fact, improve quickly if they are given a tool like this. Therefore, we have talked to people in some schools where they teach Arabic and allowed them to use our tool with each student for a period of half an hour. We had 40 Arabic students in total who use five different versions of our CALL tool. These versions are the full version (i.e. includes all facilities), animation version, synthesised version, instruction version, and null feedback version. We have divided the students into five equal groups, which means that eight students would use one type of the previous five versions. Each student had a thirty-minute session with the tool, working their way through a set of pronunciation exercises at their own pace. The exercises concentrated on sounds, which prior experience has shown, learners have difficulty with, in particular on sounds that are not differentiated in English.

The results show that the pronunciation of the students has improved using all different versions of our tool. The students have improved their pronunciation by 27% with the synthesised version, by 17% with the full version, and by 14% with the animation version. These are the highest percentages of pronunciation improvement which gives an indication that these versions (i.e. synthesised, full, animation) are the most useful ones for learners. The lowest percentage of improvement is 4% with null version. This version does not provide the student with feedback on his/her pronunciation, whether it is correct or incorrect, while other versions do. The fact that this version does not offer the user any results could be interpreted as the reason of low percentage of improvement when using the null version.

4. Conclusions

We aimed to teach non-native Arabic speakers how to sound like native speakers. Therefore, we have done a considerable amount of work to achieve our aim. Briefly, this work consists of the following; a) conducting HTK experiments on confusable sounds, b) devising a new way of carrying out an animation of the vocal tract which is done in a less computationally expensive way than similar tools such as Baldi, but it produces realistic effects, and c) performing some experiments for improving the naturalness of synthesised speech. We have integrated all this work together resulting in a CALL tool that helps learners to improve their pronunciation.

Acknowledgements. Majed Alsabaan owes his deepest gratitude to Saudi Ministry of Higher Education and Cultural Bureau in London for financial support in his PhD study. Allan Ramsay's contribution to this work was partially supported by the Qatar National Research Fund (grant NPRP 09 - 046 - 6 - 001).

References

Liu, Y., Massaro, D. W., Chen, T. H., Chan, D., & Perfetti, C. (2007). *Using visual speech for training Chinese pronunciation: An in-vivo experiment* (pp. 29-32). InSLaTE. Retrieved from http://mambo.ucsc.edu/pdf/UsingVisualSpeechforTrainingChinesePronounciation.pdf

Massaro, D. W. (2004). Symbiotic value of an embodied agent in language learning. *In System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference* (pp. 10-pp), IEEE.

Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science, 4*(4), 104-109. doi:10.1111/1467-8721.ep10772401

Massaro, D. W., & Light, J. (2003). Read my tongue movements: bimodal learning to perceive and produce non-native speech/r/and/l/. *InINTERSPEECH*.

Massaro, D. W., & Light, J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language, and Hearing Research, 47*(2), 304-320. doi:10.1044/1092-4388(2004/025)

Ouni, S., Cohen, M. M., & Massaro, D. W. (2005). Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication, 45*(2), 115-137. doi:10.1016/j.specom.2004.11.008

Ramsay, A., Alsharhan, I., & Ahmed, H. (2014). Generation of a phonetic transcription for modern standard Arabic: A knowledge-based model. *Computer Speech & Language, 28*(4), 959-978. doi:10.1016/j.csl.2014.02.005

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Povey, D., Valtchev, V., & Woodland, P. (2006). *The HTK book (for HTK version 3.4)*. Cambridge university engineering department. Retrieved from http://speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf