Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz

João P. Cabral[1], Mark Kane[1], Zeeshan Ahmed[1], Mohamed Abou-Zleikha[1],
Éva Székely[1], Amalia Zahra[1], Kalu U. Ogbureke[1], Peter Cahill[1],
Julie Carson-Berndsen[1] and Stephan Schlögl[2]

[1]School of Computer Science and Informatics, University College Dublin, Ireland
[2] School of Computer Science and Statistics, Trinity College Dublin, Ireland
joao.cabral@ucd.ie, {mark.kane, zeeshan.ahmed, mohamed.abou-zleikha, eva.szekely, amalia.zahra,
kalu}@ucdconnect.ie, peter.cahill@ucd.ie, julie.berndsen@ucd.ie, schlogls@tcd.ie

Abstract

This paper describes a prototype of a computer-assisted pronunciation training system called MySpeech. The interface of the MySpeech system is web-based and it currently enables users to practice pronunciation by listening to speech spoken by native speakers and tuning their speech production to correct any mispronunciations detected by the system. This practice exercise is facilitated in different topics and difficulty levels. An experiment was conducted in this work that combines the MySpeech service with the WebWOZ Wizard-of-Oz platform (http://www.webwoz.com), in order to improve the human-computer interaction (HCI) of the service and the feedback that it provides to the user. The employed Wizard-of-Oz method enables a human (who acts as a wizard) to give feedback to the practising user, while the user is not aware that there is another person involved in the communication. This experiment permitted to quickly test an HCI model before its implementation on the MySpeech system. It also allowed to collect input data from the wizard that can be used to improve the proposed model. Another outcome of the experiment was the preliminary evaluation of the pronunciation learning service in terms of user satisfaction, which would be difficult to conduct before integrating the HCI part.

Keywords: Pronunciation training, Wizard-of-Oz, MySpeech system

1. Introduction

The field of computer assisted systems for learning new languages has significantly grown up in recent years. For example, this evolution is reflected in the increase of commercial systems for learning pronunciation such as Carnegie Speech (www.carnegiespeech.com) and EyeSpeak (www.eyespeakenglish.com). There are also products for learning grammar and vocabulary, e.g. Rosetta Stone (www.rosettastone.com). This paper presents an early-stage pronunciation learning system and investigates how it can be used as a platform for rapid testing and development of new algorithms and pronunciation learning strategies.
Modern pronunciation tutors include several components:

- Robust and accurate pronunciation error detection module.
- Feedback generation to indicate the pronunciation errors to the user as well as ways to correct them.
- Interaction model between the learner and the system, which should be appealing and guide the student to progress in the learning process, such as spoken dialogues (Seneff et al., 2007) or games (Wik et al., 2007).
- Software interface, which needs to be easy and effective to operate.
- Pedagogical model that guides and helps the student to progress in the learning process.

Modern pronunciation training systems which use speech processing typically employ automatic speech recognition (ASR) to detect if the pronunciation of a sound or words is incorrect. Extensive research can be found in the literature about ASR methods and other speech processing techniques developed specifically for pronunciation evaluation. For example, some methods perform non-native speech adaptation (Ohkawa et al., 2009) to obtain better pronunciation error detection. Prosody is also an important aspect of pronunciation. Pitch and duration estimation methods are often used to evaluate the pronunciation, for example to detect the incorrect placement of stress in a word (Lu et al., 2010). The MySpeech system uses an ASR method for detecting pronunciation errors. It currently does not perform any adaptation to the speaker. Nevertheless, the aim of this study is the improvement of the user's interaction with the system, whereas the improvement of the pronunciation analysis component is part of future work.

Feedback to a potential user can be given in different ways. Firstly, through text by automatically generating sentences that indicate a mispronunciation and give correction instructions for solving that particular error. Secondly, by playing reference recordings spoken by native speakers, which is a way of making users perceive errors and tune their own speech production. Thirdly, by using additional visual information, such as plots of acoustic (pitch, spectrogram, etc.), phonetic and articulatory features, to help users understand their mistakes. However, users might need some experience or should receive specific training in order to be able to interpret all this different kinds of information. Consequently, an alternative approach is one that disseminates intuitive-feedback in an automated fashion, such as showing how to place and move the articulators to correct the pronunciation of a sound through an image or video of a "talking head" that displays the articulators in the speech production system. An implementation of this approach can be found in (Massaro et al., 2006). In this paper, however, we focus on two forms of feedback provided by the MySpeech system: the generation of text output and the playing of utterances spoken by a native speaker.

From a Human-Computer Interaction (HCI) perspective, the MySpeech system does not employ any advanced model such as a spoken dialogue for conversing with a user. Instead, the user must select a sentence from a list to practice the pronunciation. In addition, MySpeech allows users to choose from three different difficulty levels in order to adapt to their skills.

One current limitation of the MySpeech system is that feedback and instructions given to a user are not automatically generated. In order to develop a model for this sort of interaction we conducted an experiment using a Wizard-of-Oz (WOZ) set-up where a human imitates what the system instructions and feedback would be. The WOZ method has been used before in language learning applications. For example, it was used to study a dialogue strategy in (Ehsani et al., 2000). It was also employed to evaluate the feedback provided automatically by a speech training aid called ARTUR (Bälter et al., 2005) against the feedback provided by a phonetically trained human wizard. In this paper, WOZ is used in a different context, namely to test the HCI of the MySpeech system before developing a completely automatic interface. It also enabled us to collect data from the wizard that can be used to improve this interface.

2. The MySpeech System

2.1. Pronunciation Analysis

2.1.1. Method

The method used by the MySpeech system for analysing pronunciation variation is similar to that of (Witt and Young, 2000) with the addition of difficulty levels as described in (Kane et al., 2011). The method of the latter incorporates Broad Phonetic Groups (BPGs) to cluster similar phones, where phones that share particular characteristics such as articulatory feature information belong to the same group. This grouping of phonological units based on common phonetic features is also described in phonological theory as archiphonemes where specific features follow a markedness criteria. The categorisation of phones into BPGs allows for a difficulty level to be applied to the evaluation. For example, a difficulty level of "hard" is set by having no BPGs, hence no phonological unit is underspecified and all phonetic features that are required for that phonological unit must be present and specified. There are three difficulty levels in the MySpeech system: easy, medium and hard, whereby the easiest difficulty level includes a greater number of BPGs in comparison to the hard difficulty level. Finally, different language models are used for different levels: trigram (easy), bigram (medium) and unigram (hard). The different language models enforce that a student's pronunciation is required to have a greater acoustic capability at the hard level (unigram), whereby at an easier level acoustic variability is further tempered by the trigram language model.

The method for pronunciation evaluation can be divided into three stages which are illustrated in Figure 1. Evaluation of the student's pronunciation is based on the comparison of two phoneme strings, the known canonical phones that make-up the practice utterance and the recognised phones, similarly to (Witt and Young, 2000), which are generated in the first stage. When the known phrase is selected and attempted by the student, e.g. the phrase *see you in the morning* in Figure 1, the student's spoken phrase is force-aligned with a dictionary containing the phones for each word. This results in a file containing phones and their associated temporal information. This stage also estimates the phones for a participant's spoken utterance influenced by the difficulty level selection.

In the second stage, the canonical phones generated in stage one are temporally combined with the recognised phones, similarly to (Kane and Carson-Berndsen, 2011). The purpose of this operation is to find where the phone strings are similar at the same time. If they are similar or belong to the same BPG, the phone is assumed to be correct otherwise it is assumed to be incorrect.

Finally, a phoneme-to-grapheme conversion of the canonical phone sequence is performed in order to highlight to the user what grapheme sequences were incorrectly pronounced.

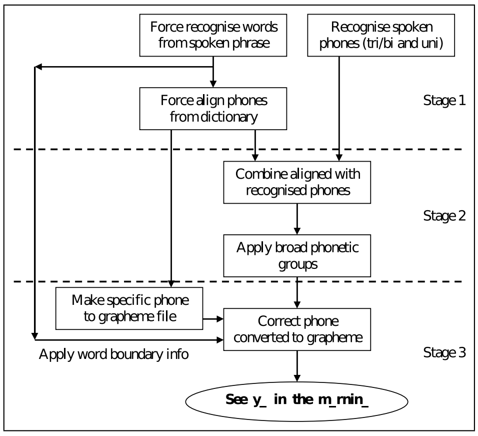

Force recognise words from spoken phrase
Recognise spoken phones (tri/bi and uni)
Force align phones from dictionary
Stage 1
Combine aligned with recognised phones
Stage 2
Apply broad phonetic groups
Make specific phone to grapheme file
Correct phone converted to grapheme
Apply word boundary info
Stage 3
See y_ in the m_rnin_

Figure 1: Block diagram of the method for phone pronunciation error detection.

2.1.2. Speech Recognition System

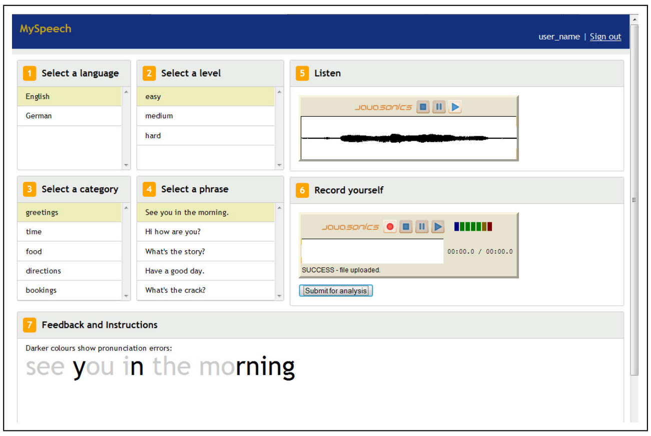
The HMM-based speech recognition system used for detection of pronunciation errors was implemented with HTK [1]. In this implementation, the HMMs consisted of five-state context-dependent triphone models that were initially calculated by cloning and re-estimating context-independent monophone models. The decoding process was implemented with a tri-/bi- or uni-gram phone model depending on the difficulty level selected by the participant and is comprised of the prompts for each phrase. The TIMIT speech corpus (Garofolo et al., 1993) was used for training

[1] http://htk.eng.cam.ac.uk/, Version 3.4.1.

the speech recogniser, consisting of read speech spoken by 630 speakers of American English.

2.2. Web Interface

Figure 2 shows a screenshot of the MySpeech web interface, which consists of several numbered panels. In panel 1 the user can select the language. The system currently supports two languages: English and German. The second panel allows the user to adapt the difficulty level (“easy”, “medium”, or “hard”). Next, there is a category panel (panel 3), so that for example, the category “greetings” can be associated with several phrases related to this domain. The different sentences are then chosen in panel 4. The audio players embedded in the interface are used by the users to listen to the selected sentence spoken by a native speaker (panel 5) and to record their own version of the same sentence and consequently submit it to the system (panel 6). Finally, the feedback panel (panel 7) shows the detected mispronunciation errors of a submitted utterance using darker colours. In this example, the submitted utterance corresponds to the sentence: *See you in the morning*.



SUCCESS - file uploaded.

Darker colours show pronunciation errors:

see you in the morning

Figure 2: Screenshot of the MySpeech web interface.

2.3. Student Database

Both the pronunciation analysis component and the web interface are connected to a database. The interface accesses the database to obtain the audio and text data for the pronunciation practice exercise. The database is also used to store data obtained from the interaction of each student with the system, including the selected sentence, recorded speech for that sentence, difficulty level and detected mispronunciations, for each pronunciation practice attempt respectively. The pronunciation analysis also requires access to the database (to obtain the recorded speech, difficulty level, etc.) and generates the information about the pronunciation errors to be stored in the database.

The aim of collecting the user’s data is to build a personalised student model that can be used to adapt the system to the user and to develop a pedagogical model (e.g. by accessing the progress of the student using this data). Currently, the system does not use such a model but the experiment conducted in this work permits to collect data from the different participants towards the development of this funcionality.

3. Experiment

3.1. Overview

An experiment was conducted in order to test a preliminary interaction model and evaluate the user satisfaction with the MySpeech system. The interaction model was simulated using WOZ. The experiment conducted was only for investigating the English pronunciation training. At the end of the experiment, participants were asked to complete a short questionnaire evaluating the system’s general usability aspects.

3.2. Interaction Model

Figure 3 shows a diagram that represents the initial model for prompting a user to practice the pronunciation of several sentences at increasing difficulty levels. The system interacts with the user through text messages shown in panel 7 of the web interface, which is shown in Figure 2 (named “Feedback and Instructions”). First, a welcome message is sent to the user to introduce the pronunciation training system. Then, the user is asked to select the difficulty level “easy” (because there is no knowledge about the language skills of the learner). She is also asked to select a category and a phrase from that category. For practising that sentence, the user is instructed to listen to the reference spoken by a native speaker, to record her own version of the same sentence and submit it to the system. After the sentence is submitted, the system performs the pronunciation analysis. The wizard has access to the results of the pronunciation analysis (generated output of the system described in Section 2.2.) and provides the appropriate textual feedback (the user only has access to this feedback, not the generated output of the system).

At the feedback step, represented by (f) in Figure 3, it is necessary to decide on the next step of the interaction. In case pronunciation errors were detected, the user is asked to repeat the same exercise for the same sentence (e) until no pronunciation errors are detected or the limit of repetitions, n_{rep}, of the same sentence is reached. Then, the user is asked to practice another sentence from the same category (d). After a user has practised a given number of sentences from the same category, n_{sent}, she is asked to select a different category (c). This iterative process is repeated for a given number of categories, n_{categ}. Once the pronunciation exercise for the level “easy” is finished, the user is asked to select the next difficulty level (“medium”) and the same iterative procedure that was conducted for the “easy” level is repeated. In this experiment, the level “hard” was never used to limit the duration of the experiment. Once the pronunciation practice exercise for the level “medium” ends, the user is informed that the experiment has finished.

The following settings were chosen for this experiment: $n_{rep} = 3$, $n_{sent} = 2$, $n_{categ} = 2$. One of the criteria for selecting these values was to limit the duration of the experiment to around 30 minutes. Another criterion for not selecting a high value of n_{rep} was to avoid a user getting frustrated or uninterested in the experiment.

3.3. The Wizard-of-Oz Setup

During the experiment a voice-over-IP system was used to give the wizard a real-time visualisation of the user’s

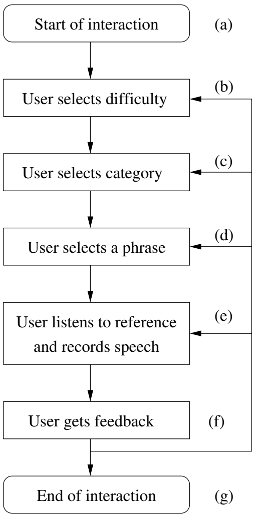
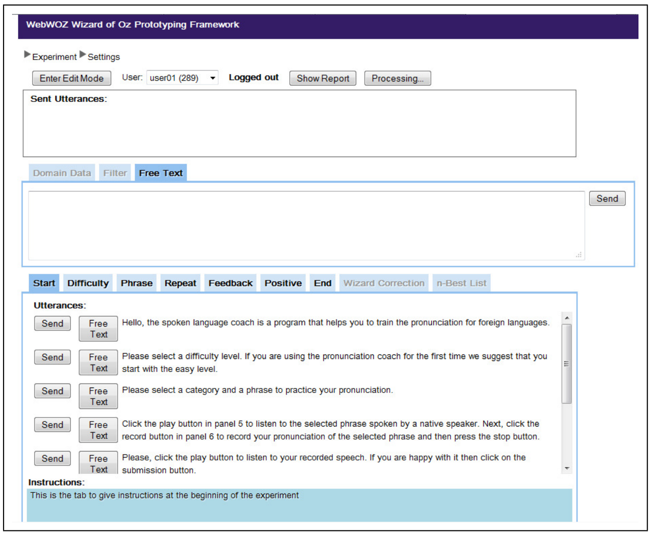


Figure 3: Block diagram of the interaction model tested in the experiment.

screen, and to transmit everything a user was saying (the user was not aware of this transmission). Furthermore, the wizard had access to the pronunciation analysis results computed by the system (displayed on a second screen). The wizard's task was to interpret a result and consequently to transform it into an appropriate textual feedback to be sent to the user. To do so a third screen, situated in front of the wizard, was dedicated to the wizard interface. A screenshot of this interface is shown in Figure 4. The wizard was instructed to not evaluate the pronunciation of the student but rather produce the appropriate feedback to the user, based on the results of the pronunciation analysis of the system.

The wizard interface as well as the intended interaction model were explained to a wizard beforehand. The interface allows for selecting predefined sentences for instructions as well as feedback. Choosing from predefined sentences as opposed to typing feedback in real-time allows for a quicker response. To decrease the time a wizard searches for an appropriate response, sentences are grouped into different panels ("start", "difficulty", "phrases", etc.). For example, the difficulty panel contains sentences that prompt the user to switch to a different difficulty level, whereas the phrases panel contains sentences that prompt her to select a different phrase. There is also a panel with encouragement messages (called "positive"), which offers a way of motivating the user. Finally, the panel called "free text" allows a wizard to input any text, or edit an already predefined sentence from one of the other panels, before sending it to the user.

The discussed interaction model was used as a guideline but not obligatory, i.e. a wizard did not necessarily need to follow it. Wizards had the freedom to alter the model as they wished (exploring the interaction space). However, most of the time they followed the guidelines and only small variations in terms of the n_{sent} and n_{categ} values were observed (perhaps as an attempt to keep the duration of an experiment close to 30 minutes). Also, offering this flexibility generated valuable feedback on the interaction model from a wizard's perspective, which can be used for further analysis and improvements. Another reason for this flexibility was to reduce a wizard's concerns about strictly following the guidelines, which otherwise could have caused an increased delay in response time. Since an important factor for the usability of the overall system is the time it takes for a user to get feedback, wizards were asked to respond as fast as possible.



Hello, the spoken language coach is a program that helps you to train the pronunciation for foreign languages.

Please select a difficulty level. If you are using the pronunciation coach for the first time we suggest that you start with the easy level.

Please select a category and a phrase to practice your pronunciation.

Click the play button in panel 5 to listen to the selected phrase spoken by a native speaker. Next, click the record button in panel 6 to record your pronunciation of the selected phrase and then press the stop button.

Please, click the play button to listen to your recorded speech. If you are happy with it then click on the submission button.

This is the tab to give instructions at the beginning of the experiment

Figure 4: Screenshot of the WOZ web interface.

3.4. Participants

Ten postgraduate students participated in the experiment as users of the MySpeech system. They were all non-native English speakers. Six different wizards participated in the simulation of the interaction through using the WOZ platform. They were all part of the team who developed MySpeech and therefore familiar with the intended system behaviour.

4. Results

4.1. User's Satisfaction

Table 1 shows the overall results obtained from the questionnaire.

From the questionnaires, examples of relevant comments given by participants are listed below:

- it would be better if the system was faster in the response.
- sometimes the system was too strict in the pronunciation error detection.
- sentences could be different between the easy and medium levels.
- the system could give more feedback on how to correct pronunciation errors.
- the user interface was simple and clear.

Questions	Answers
Please rate the instructions given by the spoken language coach: 0 (not enough) to 5 (very good)	Mean: 4
Please indicate if you would use the spoken language coach to learn common phrases in a foreign language (yes/no)	"Yes": 60%
Please indicate, how often the spoken language coach helped you understand and correct pronunciation errors (never/sometimes/many times/always)	"Sometimes": 80% "Many Times": 20%
What is your general appreciation of the service provided by the spoken language coach? (poor/good/very/good)	"Good": 80% "Very good": 20%
Would you use the spoken language coach again for improving your pronunciation in foreign languages?	"Yes": 70% "No": 20%; "Maybe": 10%
Was it fun to use the spoken language coach? (yes/no/sometimes)	"Yes": 50% "No": 20%; "Sometimes": 30%

Table 1: Results obtained from the answers of the users to the questionnaire.

- it would be easier if the system automatically selected the sentences.

In general, the MySpeech service provided a good pronunciation training service for non-native English speakers. The results indicated that the interaction model tested in the experiment to give instructions and guide the user was effective, as the participants generally rated the instructions given by the system as good.

A weak point of the system is the actual feedback. It needs to be improved as for the user to better understand and correct the pronunciation errors. Also, some participants stated that sometimes the system was not precise enough in indicating the pronunciation errors. This could be improved by generating more detailed feedback messages about the pronunciation errors. In this protoype version of MySpeech, users correct pronunictaion errors by listening to the native and their own recorded speech and tunning their speech production. However, the comments indicated that users would prefer the system to give instructions for helping them to correct the pronunciation errors. This component could be developed in the future, for example by using a talking head and specific pronunciation correction instructions.

The participants in the experiment were generally happy with the system and stated that they would use it again in the future. However, there seems to be room for an improved interaction model that is more fun and engaging for the user. A possible solution might be the integration of the MySpeech system into an online game.

4.2. Wizard-of-Oz and Student's Input Data

The data obtained from the users corresponded to 465 attempts (submitted utterances for pronunciation evaluation), which included: recorded speech from the user, the individual pronunciation errors (at phone level) for each attempt and the selected difficulty level. This data can be used in the future to create student profiles for developing a pedagogical model.

The total number of messages i.e. instructions/feedback sent by the wizards to the MySpeech interface was 512 (on average 51 messages per session). Table 2 shows the distributions of the total number of messages used by the wizards relatively to the different types of messages that were part of the WOZ interface: corrective feedback, positive feedback, and instructions. Tables 4 to 5 show the top three messages that were sent by the wizards using the interface, for the three message types. The corrective feedback sentences are for indicating where the individual mispronunciation errors are within the sentence or word. Meanwhile, positive feedback messages relate to positive pronunciation assessment, i.e. whether user pronounced correctly or improved pronunciation of the sentence, and encouragement messages that are important for the pedagogy aspect. The third type of messages are mainly intended to guide the user through the steps of the HCI model shown in Figure 3.

Message Type	Panel	Total	New	Edited
Instructions	23	276	7.2%	0%
Corrective Feedback	12	125	8.8%	88%
Positive Feedback	8	111	1.8%	0%

Table 2: Quantitative analysis of the type of messages used by the wizards, for the main message panels of the WOZ platform. The second column contains the number of messages of each panel, the third column represents the total number of messages sent by the wizards, the fourth column respresents the rate of the total messages that were typed by the wizards without using any messages from the panels and the last column indicates the rate of the total messages that were obtained from the panels and edited by the wizards before they sent them to the users.

Rate	Message
85.7%	You mispronounced the last part of the word
26.1%	Please try to emphasize
25.2%	You mispronounced the word

Table 3: Most frequent messages from the corrective feedback panel sent by the wizards.

The relatively low rates (under 10%) of new messages typed by the wizard shown in Table 2 indicate that the mes-

Rate	Message
13.7 %	Next, try a different phrase from the same category
13.4%	Try listening to the same reference phrase again and record your pronunciation
9.42%	Next, try a different phrase from a new category

Table 5: Most frequent messages from the instructions feedback panels sent by the wizards.

Rate	Message
21%	Perfect, you pronounced the phrase correctly
15%	You are showing some improvement
14%	You are almost there

Table 4: Most frequent messages from the positive feedback panel sent by the wizards.

sages selected for the HCI model used in this experiment were appropriate and represented well the type of messages that the wizard needed to use.

These results also indicate that positive feedback can be provided to the users using a lower number of messages than for corrective feedback. In fact, eight messages for positive feedback appeared to be sufficient in this experiment as the rate of new messages typed by the wizards for this case was very small (1.8%). A higher variability of this type of messages could contribute to a more "human" or "realistic" user's experience while interacting with the system, as humans tend not to repeat exactly the same sentences. However, in this experiment few messages with similar meaning were used in order to facilitate a quick selection of messages by the wizard and because with the WOZ method the users should have the illusion that they are talking to a computer system rather than a person.

Table 2 also shows that there was a more intense activity from the wizards for sending corrective feedback messages as the rate of edited messages was higher for this type. An example of this type of message was "You mispronounced the words: *you and rooms*", in which the emphasised text indicates the part of the sentence typed by the wizard. From the results, it appears that the set of these sentences was adequate for the wizard to provide feedback to the user as a low rate of new sentences were typed by the wizards. However, the system needs to be further developed to automatically generate this type of sentences because they currently require human intervention.

5. Conclusion

This paper describes an early stage pronunciation training system for non-native languages, called MySpeech. The user interacts with the system through a web-based interface that permits different operations:

- Selection of sentences from different topics by the user.
- Selection of a difficulty level by the user.
- Playback of selected sentence spoken by a native speaker.
- Recording and playback of user's speech.
- Provide feedback and instructions to the user by the system.

In order to develop an interaction model to guide the user through the learning process and evaluate the preliminary version of the system in terms of user satisfaction, an experiment was conducted employing the WOZ method. The wizard guided the user through the selection of sentences and difficulty level. Another function of the wizard was to provide feedback to the user based on the results of the pronunciation analysis generated by the MySpeech system. Results showed that the MySpeech system provided a good service for non-native speakers to train their English pronunciation. Also the interaction model of the system performed well.

The next step is to implement the interaction model in the system for automatic generation of instructions and feedback messages regarding pronunciation assessment. The data obtained from the wizard is also being studied to obtain a better set of feedback messages. Finally, future work also includes the incorporation of additional modalities for providing feedback, such as a talking head, in order to show more details about pronunciation errors and suggest ways for correcting those errors.

6. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin and Trinity College Dublin. The opinions, findings and conclusions, recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

7. References

O. Bälter, O. Engwall, A. Öster, and H. Kjellström. 2005. Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction. In *Proc. of the Seventh International ACM SIGACCESS Conference on Computers and Accessibility*, pages 36–43, Baltimore.

F. Ehsani, J. Bernstein, and A. Najmi. 2000. An interactive dialog system for learning Japanese. *Speech Communication*, 30(2-3):167–178.

J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, 1993. *The DARPA TIMIT Acoustic-phonetic continuous speech corpus CDROM*.

M. Kane and J. Carson-Berndsen. 2011. Multiple source phoneme recognition aided by articulatory features. In *Proc. Springer-Verlag Lect. Notes in Comp. Sci. (IEA/AIE)*, pages 426–435, Syracuse, NY.

M. Kane, J. P. Cabral, A. Zahra, and J. Carson-Berndsen. 2011. Introducing difficulty-levels in pronunciation learning. In *Proc. of SLaTE*, pages 37–40, Italy.

J. Lu, R. Wang, L.C. De Silva, Y. Gao, and J. Liu. 2010. CASTLE: a computer-assisted stress teaching and learning environment for learners of English as a second language. In *Proc. INTERSPEECH*, pages 606–609, Japan.

D.W. Massaro, Y. Liu, T.H. Chen, and C. Perfetti. 2006. A multilingual embodied conversational agent for tutoring speech and language learning. In *Proc. INTERSPEECH*, pages 825–828, Pittsbourgh, PA.

Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito, A., and S. Makino. 2009. A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning systems. *Speech Communication*, 51(10):875–882.

S. Seneff, C. Wang, and C. Chao. 2007. Spoken dialogue systems for language learning. In *Proc. NAACL HLT07*, Rochester, NY.

P. Wik, A. Hjalmarson, and J. Brusk. 2007. DEAL A Serious Game For CALL Practicing Conversational Skills In The Trade Domain. In *Proc. of SLATE*.

S. M. Witt and S. J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108.