John Sloan, Beate Luo*, Julie Carson-Berndsen

University College Dublin, Dublin, Ireland
*Feng Chia University, Taichung, Taiwan

john.sloan.1@ucdconnect.ie, beate@fcu.edu.tw, julie.berndsen@ucd.ie

Emotional response language education: a first 'in-the-wild' evaluation

Bio data

John Sloan is a PhD candidate in the School of Computer Science at University College Dublin under the supervision of Professor Julie Carson-Berndsen. His research is centered on the development and testing of a personalised, e-learning platform where feedback to language learners is provided through the facial expressions of an animated avatar. His background includes teaching ESL and an MA in Linguistics from UCD in 2016.

Beate Luo works as an associate professor at the Department of Foreign Languages and Literature of Feng Chia University in Taichung, Taiwan. Her research interests include CALL/MALL, vocabulary acquisition, and pronunciation training.

Julie Carson-Berndsen is a Professor in the UCD School of Computer Science where her research group has developed phonetic-feature based approaches to speech recognition and expressive speech synthesis systems. Her current research focusses on spoken language analytics using data-driven syntagmatic and paradigmatic similarities to approximate native speaker intuitions for non-native learners and virtual agents.

Abstract

This paper reports on the development and testing of Version 3 of the Emotional Response Language Education (ERLE) e-learning platform. An 'in-the-wild', heuristic user evaluation with five English as a Foreign Language students from Feng Chia University in Taiwan and one native English speaker in Ireland was performed over three months, with feedback from students informing changes and improvements. The primary goal of the study was to assess the robustness and reliability of a newly integrated speech recognition system to the ERLE platform. The feedback garnered led to the introduction of a tutorial prior to the initial class, a redesign of the buttons and presentation of the ASR output, and an animated response to loud input which causes difficulty for the ASR system. The improved system has since been implemented as a complimentary aid to a first-year English speaking and listening course at the same university in a larger, longitudinal study.

Introduction

ERLE is an e-learning platform which enables English language learners anywhere in the world to interact with a native speaker through the medium of an animated avatar on the erle.ucd.ie website. By employing 3D WebGL technology, the platform was developed to provide learners with an engaging and immersive virtual interaction (Sloan & Carson-Berndsen 2018). A human-like avatar displays live, native-speaker feedback primarily through change in facial expressions and gaze (see figures 1&2). This form of feedback affords the learner an opportunity to receive instant, accurate and consistent evaluation on their production while maintaining a familiar learner-teacher interaction, and has been shown to prompt language learners into reflecting on their production and altering the complexity of further utterances in response (Sloan & Carson-Berndsen 2017). The pedagogical method of providing learners with explicit, meta-linguistic information on all errors is based on both the first authors experience as an ESL teacher and numerous empirical studies demonstrating the effectiveness of this method (Li, 2010; Lyster et al., 2013). The interaction initiated after an ill-formed utterance (shown bottom-right in Figure 2), where the learner can choose to try again, ask to see what's wrong (Figure 3) or attempt another sentence follows from task-based studies demonstrating improved vocabulary and grammar retention when interaction is allowed (Gass & Veronis, 1994; Mackey, 1999)

Figure 1: Main ERLE user interface with recording and listening buttons on the right.

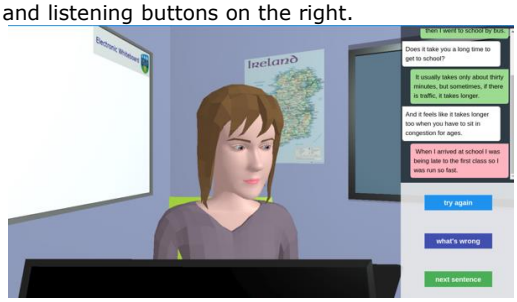

Figure 2: Expression and interaction buttons after ill-formed learner sentence

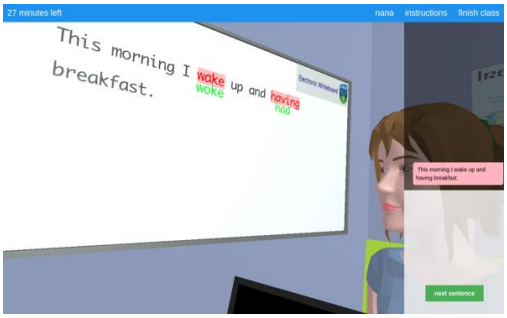

Figure 3: Error and Correction Displayed

The user interface to an application is crucial for affording the users the ability to perform the tasks for which it has been designed. This is of particular importance when providing language learners with explicit corrective feedback on language production. They are taking part in a cognitively demanding task where correction can lead to embarassment, frustration and anxiety. Therefore, creating an easy-to-understand, highly usable interface is essential to minimise non-linguistic related errors (i.e. pressing the wrong button) which could lead to a negative experience. Two widely used forms of interface testing are user testing and heuristic evaluation. User testing usually involves scenario-based tasks, while heuristic evaluation is based on the user exploring and evaluating the interface on their own terms (Tan et. al, 2009). The major advantages of heuristic testing over user testing that it is relatively quick to implement and can be used early in the design process. Nielsen & Molich (1990) demonstrated that such a method, whereby users interact with a platform and informally give their opinion on what is good and bad about the design, identified 55-90% of usability problems on the four different interfaces they tested. This was achieved with as few as five evaluators. Jeffries et al. (1991) showed heuristic analysis could discover three times more errors than standard user testing. The success of heuristic evaluation along with its flexibility and ease of implementation made it the best choice for this study.

The interaction with the avatar on ERLE was designed with the aim of informing the learner, with the least amount of written and spoken instruction, how to use the interface. This is in line with the principle of minimising the users' cognitive load through human-centered design (Oviatt, 2006). Consideration for reducing this load is of even higher concern when the end users are intermediate-level ESL learners who are already preparing for a demanding task – speaking in English. Where possible, easy to understand signifiers (Norman, 1988), e.g. microphone and square stop symbols, were included to be understandable by users regardless of first language. The avatar is programmed to begin each class with a greeting, then ask the learner how they feel and what they would like to talk about. The learner is then instructed "Please begin when you are ready", and input buttons appear on the right-hand side (Figure 1). From this point, the learner is able to control their input and interaction with the avatar for the duration of the class – 30 minutes – after which, class automatically ends. Users are expected to be able to navigate the website, begin and maintain an interaction with the avatar using the speech recognition for the full 30 minutes. Any deviation from this is to be deemed a failure of the interface design and warrant a potential redesign.

Method

A remote heuristic user evaluation study with five EFL students from Feng Chia University in Taiwan was carried out over three months. The five evaluators were first given access to the e-learning system via the web for two weeks, where they took classes and provided feedback on their experience. This feedback was used to inform a major redesign of the interface to improve the user experience. After one month of development, the same evaluators were given access to the platform and performed another round of evaluation. This was used to locate problems and make final changes and improvements to the interface to create an easy-to-use and robust system for future users.

To gather detailed, honest user feedback which would maximise the benefit to the system, it was important to have regular and close contact with users who evaluated the platform over a number of weeks. Heuristic user evaluation requires that evaluators are free to access and interact with the interface as they see fit. The physical distance and time difference between Taiwan and Ireland (8 hours) were significant barriers to overcome in this form of study. The use of a mobile instant messaging (IM) service was necessary to allow for contact between developer and evaluators when the evaluators wished to use the platform. 'Remind', was used for the initial two weeks of testing, then 'Line' was brought in as a replacement due to teacher and student preferences.

The researcher posted initial instructions on how to log on and enter the first class in a group chat. Later, when evaluators used the ERLE platform, the researcher was concurrently in contact with the students before, during and after the class on the

messenger platform. Evaluators were invited to post messages whenever they felt something was wrong, or they were not sure what to do (see figures 6, 8 & 9). After class, the evaluators were invited to give their opinion on the class. No evaluators were required to continue doing more classes if they did not wish, as drop-out is an important indicator of engagement.

Concerns over the collection and storing of user data increased with the introduction of speech recognition to the ERLE platform. As voice recordings are potentially identifiable, it was of paramount importance that the evaluators were fully aware as to how their data is stored, and their right to obtain and delete it. As such, the information sheet and consent form for the study were presented to the evaluators in both simple English and a translation of that into their native Chinese. Each point of consent had to be individually clicked to be confirmed before the evaluators were allowed to begin participation. These included statements on the storing of all audio and text entered. All data is stored on a private, secure server which only the researchers and evaluators themselves can access.

Results

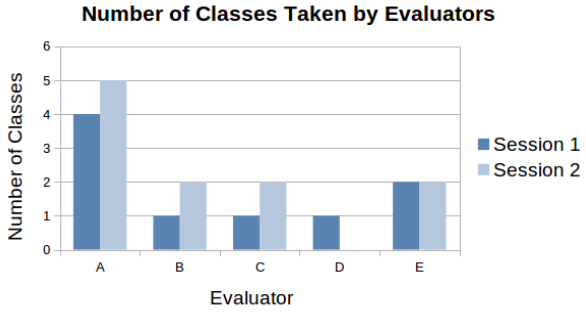

Figure 4: Graph showing classes taken by evaluators A, B, C, D & E over the two sessions

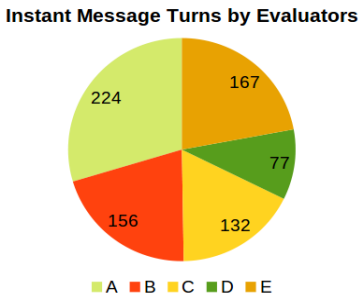

Figure 5: Chart displating number of Instant Messenger turns completed by each evaluator over both sessions

A total of 20, 30-minute classes were taken by the five evaluators over two sessions. 9 classes were taken in the first 2-week session, the remaining 11 being completed in Session 2 after the month-long redesign (Figure 4). The total number of Instant Messaging 'turns' was 756 (One turn meaning a message or series of messages from an evaluator followed by one or more in response from the researcher, e.g. Figure 6 shows 2 turns), with 57 of those including screenshots from the evaluators (Figure 5).

Feedback from evaluators falls into two distinct categories. Queries and statements of uncertainty on how to navigate the interface and interact with the avatar were prominent in the first class. Figure 6 shows an exchange from one student which demonstrated that the minimal initial instructions were not sufficient to inform all users as to how to begin the interaction. In later classes, evaluators commented more on aspects of the interface and interaction which appeared erroneous or problematic, providing screenshots of the interface and the issue where necessary. Examples of three pieces of feedback and the changes implemented are detailed in the following section.

Feedback & Changes

Three major changes were implemented to the platform following the feedback provided from the users during the three-month testing period: a tutorial was introduced prior to the initial class; the ASR output to students and buttons for manipulating it were significantly changed; a means of dealing with background noise and loud input by using the visible physiological response of the avatar was put in place. These changes, and the conversations on messenger services which brought them about, are detailed below.

Change 1 – Tutorial

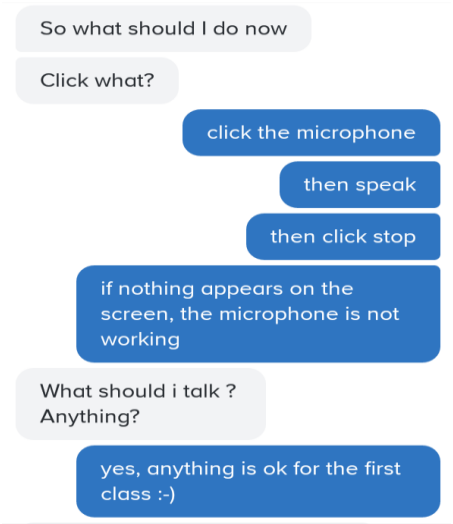

Figure 6: IM example with evaluator B

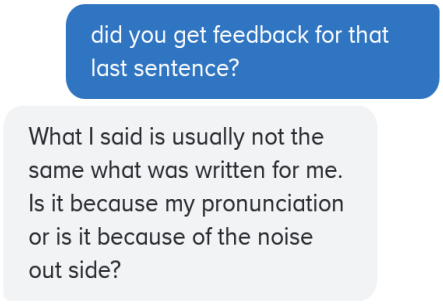

Figure 8: IM example with evaluator A - pronunciation query

The ERLE interface had been designed to be easy and intuitive for users of any language. Icons for the buttons were chosen to be universally recognized (e.g. the microphone icon seen in the blue button in Figure 1). Three students were able to navigate and work out the correct usage of the platform with minimal to no assistance, but two required explicit instructions over IM as to how to engage with the avatar. The two IM turns in Figure 6 showed that evaluator B did not understand from the avatar's instructions and symbols in the microphone buttons how they were expected to proceed. This included both the topic of conversation and using the functionality of the site.

The avatar, Tia, at the beginning of a class, asks the student what they would like to talk about, and the student must choose a topic to proceed. The microphone button then appears, indicating that it should be clicked and it is the student's turn to speak. There are many reasons a user may not know how to proceed in this case, including having an insufficient English ability to fully comprehend the avatar's instructions to even simply not paying attention to the introduction. Therefore, it was decided that a short tutorial (Figure 7) which included examples and practice of the main scenarios encountered in a class would be beneficial. Each step of the tutorial must be carried out successfully before the user can enter their first real class.

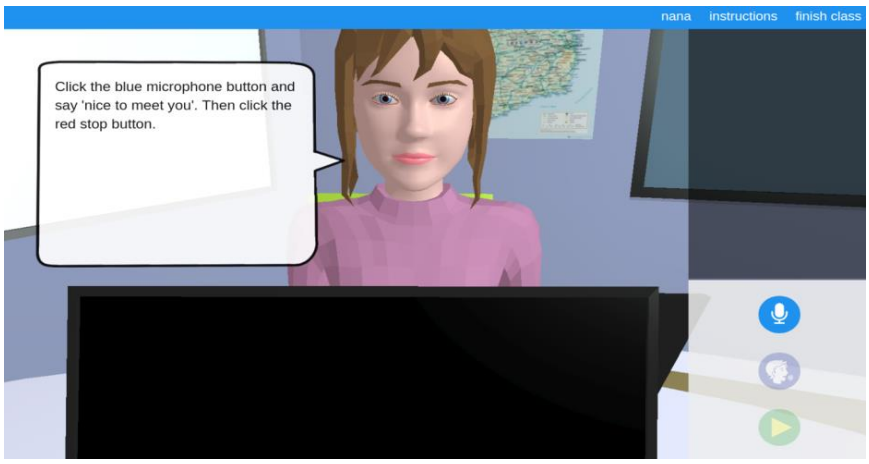

Figure 7: Example from the tutorial added in the redesign explicitly stating how to interact with the avatar via speech

Change 2 – ASR Output Representation

The speech-to-text system frequently does not provide output which matches the user's intended utterance. This occurs most frequently due to non-native pronunciation, e.g. the user intends to say, "I think that...", but the ASR outputs, "I seen that...", due to the user erroneously pronouncing [θ] as [s]. However, non-target output can also be due to a number of other factors, including microphone quality and volume, background noise and

the language model underpinning the ASR not containing grammatically erroneous word sequences which non-native speakers utter.

The ASR used by ERLE is the WEB Speech API of Google's Chrome browser. Users click the microphone button, say a sentence and then click the stop button. This is transcribed to text by the ASR and displayed to the user. If the text displays the desired target output, the user may send this sentence to the avatar to receive grammatical and lexical feedback. If the ASR output is different from the user's intended utterance, then they have the option of trying to record again or typing and correcting the section which differs. Three of the evaluators made comments that demonstrated their uncertainty over the ASR output. Figure 8 shows evaluator A questioning why the output is usually different. A major problem which emerged was the evaluators growing mistrust of the ASR system as it failed to output their desired target despite repeated tries.

To solve the problem of user frustration due to repeated non-target ASR output, two changes were made during the redesign. First, a section of the new tutorial explaining why the ASR may provide different output was included. Second, it was decided to display to users up to three ASR outcomes for each spoken utterance. The user can now click through the 3 transcriptions to see what possible alternatives exist and which words and phrases are confusing the system. In Session 2, this appeared to improve the evaluators understanding of both the workings of the ASR system, and their own pronunciation errors.

Change 3 – Interference Signal

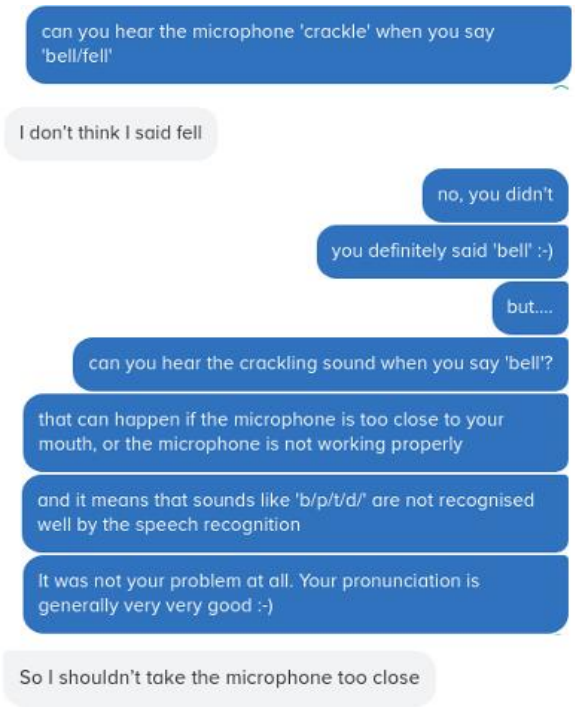

Figure 9: IM example with evaluator A - microphone proximity problem

As this is both an 'in-the-wild' and a remote heuristic user evaluation study, the researcher had no control over the conditions under which the evaluators used the platform. As a result, each evaluator had different conditions from which they provided speech input. These differences included the type of microphone being used, the proximity of the microphone to the user's mouth, the input volume and background noise.
Checks on the microphone volume and instructions to minimise background noise were included in the tutorial. However, the adverse effects of microphone proximity on the ASR output was difficult to convey to the evaluators. This is demonstrated in the IM turns shown in Figure 9. Evaluator A said "bell", but due to the close proximity of their headset microphone, the aspiration from the plosive [b] next to the microphone registered more energy. This was recognised as a fricative, which, the language model predicted as an 'f' in 'fell'. Evaluator A, upon listening to their own recording of the attempted 'bell', could clearly hear the 'b', not 'f'. This caused confusion and a loss of trust in the ASR.

Explaining the problems associated with background noise and microphone volume to users is relatively straightforward and achievable with the tutorial. However, doing the same for problems related to individual phoneme realisations and the workings of language models is more complex. Therefore, four solutions have been put in place which improved the quality of audio recordings in Session 2. These are a visible volume bar when the user is speaking, a negative expressive reaction from the avatar when the input volume passes

an upper limit, an interference sound played back to the user, and a verbal signal from the avatar (see figure 10). These guide the users toward maintaining a minimum distance between the microphone and mouth to provide input which can be more accurately transcribed by the ASR, all while minimising the cognitive load on the user.

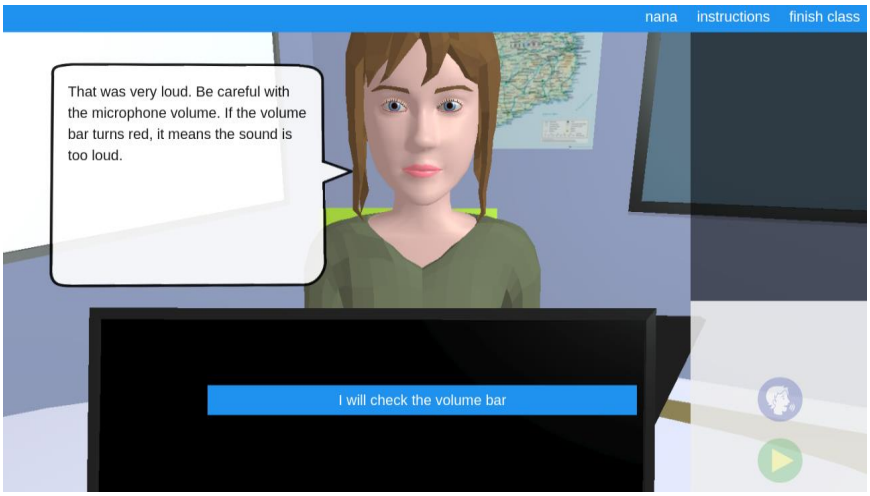

Figure 10: Example feedback from the avatar when the spoken user input is too loud

Conclusion

With 5 users, 10 hours of classes and 756 IM turns, the ERLE platform was tested, feedback gathered, a redesign implemented, and a second session of testing and feedback carried out. Following the heuristic user evaluation model proved to be successful in identifying major problems users had with the platform. It allowed the evaluators freedom to test when and how they wished, and the researcher the flexibility to make, implement and test changes rapidly. Three major changes were implemented along with a larger number of minor ones and these have improved the robustness and usability of the platform. The improved platform is now being used successfully by a larger number of students in a longitudinal study to test the effectiveness of the pedagogical method based on feedback provided through expression.

References

Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in second language acquisition*, *16*(3), 283-302.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991, April). User interface evaluation in the real world: a comparison of four techniques. In *CHI* (Vol. 91, pp. 119-124).

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309-365.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language teaching*, *46*(1), 1-40.

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in second language acquisition*, *21*(4), 557-587.

Nielsen, J., & Molich, R. (1990, March). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249-256). ACM.

Norman, D. A. (1988). *The psychology of everyday things* (Vol. 5). New York: Basic books.

Oviatt, S. (2006, October). Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia* (pp. 871-880). ACM.

Sloan, J., & Carson-Berndsen, J. (2017). Was it something I said? Facial Expressions in Language Learning. In *SLaTE* (pp. 1-6).

Sloan, J., & Carson-Berndsen, J. (2018). Expressive Data: A learner corpus with emotion. *Proceedings of Computer Assisted Language Learning Conference XIX*.

Tan, W. S., Liu, D., & Bishu, R. (2009). Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics*, *39*(4), 621-627.