Evaluating automatic speech recognition-based language learning systems : a case study

Evaluating automatic speech recognition-based language learning systems: a case study

Joost van Doremalen[1], Lou Boves[1], Jozef Colpaert[2], Catia Cucchiarini[1], Helmer Strik[1]

[1] Centre for Language and Speech Technology, Radboud University Nijmegen
Erasmusplein 1, 6525 HT Nijmegen, The Netherlands.

[2] Institute for Education and Information Sciences, Universiteit Antwerpen, Belgium.

E-mail: j.vandoremalen@let.ru.nl

Evaluating automatic speech recognition-based language learning systems: a case study

The purpose of this research was to evaluate a prototype of an automatic speech recognition (ASR)-based language learning system that provides feedback on different aspects of speaking performance (pronunciation, morphology and syntax) to students of Dutch as a second language. We carried out usability reviews, expert reviews and user tests to gain insight into the potential of this prototype and the possible ways in which it could be further adapted or improved, with a view to developing specific language learning products. The evaluation revealed that domain experts and users (teachers and students) are generally positive about the system and intend to use it if they get the opportunity. In addition, recommendations have been made which range from specific changes and additions to the system to more general statements about the pedagogical and technological issues involved. These recommendations can be useful to improve this prototype and to develop other ASR-based systems, which can be deployed either as language courseware or as research tools to investigate design hypotheses and language acquisition processes.

Keywords: computer-assisted language learning, automatic speech recognition, computer-assisted pronunciation training

1. Introduction

Recent views on second language (L2) acquisition emphasize the importance of usage-based learning and skill-specific practice (DeKeyser & Sokalski, 2007; Ellis & Larsen-Freeman, 2009): for learners to speak the L2 fluently and accurately, they should practice speaking it and receive appropriate feedback. Unfortunately, in teacher-fronted lessons there is generally not enough time for sufficient practice and feedback on speaking performance, while traditional language lab tools usually do not provide the feedback required. This is in line with findings by Dlaska & Krekeler (2008) which show that it is difficult for L2 learners to evaluate their own pronunciation.

Against this background, various systems have been developed that employ Automatic Speech Recognition (ASR) technology to provide practice and feedback for L2 speaking, such as FLUENCY (Eskenazi, 1996), EduSpeak (Franco et al., 2000), Tell me More (www.tellmemore.com), the Tactical Language Training System (Johnson et al., 2004), the SPELL system (Morton & Jack, 2010), Carnegie Speech NativeAccent (Eskenazi et al., 2007), Saybot (Chevalier, 2007; www.saybot.com), and Rosetta Stone (www.rosettastone.com). A recent overview of ASR-based commercial systems for L2 pronunciation is provided by Witt (2012).

Many of these systems, however, do not contain important and required features of feedback on L2 pronunciation, such as immediate, detailed feedback on individual segments in connected speech. Recent reviews of technologies for language learning show that there is little understanding of the role of ASR technology in computer assisted language learning and its potential contributions (Golonka et al., 2012; Steel & Levy, 2013). Furthermore, systems that address grammar skills like morphology and syntax generally do not support spoken interaction (Bodnar et al., 2011). In addition, most of these systems address English, while fewer products are available for languages other than English, for instance Dutch.

It was in this context that a new project aimed at realizing and testing a prototype of an ASR-based Computer-Assisted Language Learning (CALL) system for Dutch L2 (DL2) speaking was started. For DL2 there are some commercial systems (Tell me More, Rosetta Stone), but so far there was no open system, that is, a system for which there is a clear explanation of how it works and how it performs that could be used for research and development in ASR-based CALL. The opportunity to realize a system of this kind arose within the framework of a speech technology research programme funded by the Dutch and Flemish governments (STEVIN). Since tool development and CALL data were among the priorities of this programme, a project, *Development and Integration of Speech technology*

into COurseware for language learning (DISCO) was started with the aim of developing an ASR-based system. This system had to automatically detect pronunciation errors (mispronunciations of speech sounds) and grammar (morphology and syntax) errors in DL2 speaking and to generate appropriate, detailed feedback on the errors identified. The embedding of DISCO in a government-funded programme partly explains why its aim was not to realize a commercial product for DL2 speaking but to test ASR technology for its added value for language learning and teaching.

In the course of the project, various experiments have been conducted to test the various technology components. An important aspect in this kind of research is ASR performance, since it is known that automatic recognition of L2 speech can be problematic (Benzeghiba et al. 2007). In previous papers we showed that the performance of speech recognition was satisfactory even for such low proficient speakers as our target group (van Doremalen, Cucchiarini & Strik, 2010) and that error detection was sufficiently accurate (Cucchiarini, van Doremalen & Strik, 2012). Additional factors such as general design, user interface and interaction patterns have received less attention in the literature on ASR-based CALL, while we hypothesize that learner analytics can reveal important information about the learning process. For this reason we carried out usability reviews, expert reviews and user tests to gain insight into the potential of this prototype and the possible ways in which it could be further adapted or improved.

In this paper we report on these latter types of evaluations, while for evaluations of the technology components we refer to (van Doremalen, 2014). The current paper is organized as follows. We first introduce the DISCO system (Section 2) and discuss the evaluation methods in Section 3. In Section 4 we present the results of these evaluations, and in Section 5 we discuss these results and present future perspectives. Conclusions are drawn in Section 6.

2. DISCO system overview

In this section we will present background information on the DISCO system. We first discuss the design of the system in Section 2.1. In Section 2.2 we present an interaction walkthrough of the system.

2.1 System design

The design of the DISCO system is based on three stages: conceptualization, specification and prototyping.

The first stage, the *conceptualization* of the DISCO system (Colpaert, 2013), is based on the Distributed Design model, an educational engineering approach proposed by Colpaert (2014) and currently still under empirical and theoretical validation. The model is based on a number of striking hypotheses, amongst which

- *The Ecological Paradigm Shift:* No technology carries an inherent, measurable and generalizable effect on learning. This effect can only come from the entire learning environment as an ecology. The role of any educational artefact, and of any technology like ASR or CAPT, is to contribute to the global effect of this learning environment as a piece of a puzzle.
- *The Process-oriented Paradigm Shift:* The targeted learning effect is proportional to the designedness of the learning environment. This means the extent to which it has been designed in a methodological and systematic way. The reasoning behind the way an environment has been designed is far more important than the features of the eventual product, which should by definition be different in every single learning context.

- *The Psychological Paradigm Shift* states that in most learning contexts it is counterproductive to focus exclusively or too intensively on pedagogical goals (Colpaert, 2010). In order to create willingness and acceptance in the learners' mind first, it is more effective to focus on personal goals first. The problem with personal goals is that they are difficult to elicit, and a special technique is needed to identify them.

Even if the initial goal of the DISCO project was (only) to test ASR-technology in CALL, and not to develop a full-fledged market ready product, we decided to adopt the proposed design approach as if we were developing the system for the entire population of DL2 learners in Flanders and the Netherlands. The reasoning behind this choice was the following: It was possible that CALL-integrated ASR-technology performs remarkably well from a technological and even linguistic-didactic point of view, but that problems would arise on the level of design: lack of acceptability of the proposed interface, labour-intensive content development or poor sustainability.

We started by identifying the *pedagogical* goals, which were formulated as follows: (1) to develop exercises and automatic feedback moves that help improve grammar and pronunciation in high-educated DL2 speakers at A2 CEFR level, (2) to integrate the pronunciation and grammar exercises in a communicatively-oriented method and (3) to provide remedial exercises which help the DL2 learner improve pronunciation and grammar on specific linguistic structures.

To establish the *personal* goals, we needed to conduct a number of specific focus groups and in-depth interviews (Strik et al., 2009). The most important personal goals were: (1) DL2 learners want to practice in a safe environment which helps them to gradually and repeatedly improve their pronunciation and grammar skills, (2) DL2 learners want to receive tailored feedback when practicing communicative skills in general and pronunciation

specifically, and (3) DL2 learners do not like an exaggerated focus on what they perceive as 'back to school' or 'adapt and integrate' Ought-to Selves, but they see a natural interaction with local natives as a visualization of the roadmap to their IDEAL Self (Dörnyei and Ushioda, 2009)

In order to try to find a working compromise between these pedagogical and personal goals, we decided to limit our design space to closed response conversation simulation and interactive participatory drama (Hubbard, 2002), a genre in which learners play an active role in a pre-programmed scenario by interacting with computerized characters or "agents". The simulation of real-world conversation is closed: students choose the words they use in their responses from the screen. In most turns, students can choose between responses that influence the course of the dialog, which grants them some amount of conversational freedom.

As a second stage in design, we specified a communicative CALL application that stimulates DL2 learners to produce speech. More importantly, the framework allows us to circumvent most of the limitations of today's ASR technology, which are primarily related to the impossibility of handling unpredictable, spontaneous speech from L2 learners. For this reason, strategies aimed at constraining the learner's output to make the speech more predictable are often applied in this context (van Doremalen, Cucchiarini & Strik, 2011).

The learning process in the program starts with conversation simulation (a dialog). Based on the type of errors the students make, they are then offered remedial exercises, which are exercises that focus on specific speech sounds or syntactic and morphological structures without a conversational context. The feedback strategy is immediate corrective feedback visually implemented through highlighting, which puts the conversation on hold and focuses on the errors. Initially, three dialogs were developed. The topics of these dialogs are (1) travelling by train, (2) choosing a hobby/course and (3) buying a DVD player.

Each of these dialogs can be conducted in three different modes or exercise types:

(1) **Pronunciation exercises:** Pronunciation exercises consist of reading one of the response options offered by the system, so that the quality of the speech sounds pronounced by the learner can be evaluated. The choice of speech sound to be addressed was based on results of previous research (Neri et al. 2006).
(2) **Morphology exercises:** We opted for a multiple-choice approach. Within the response options, morphological variants are presented on the screen. For example, for personal pronouns: "Hoe gaat het met (hij/jou/wij)?" ("How are (he/you/we)?") and for verb inflections: "Hoe (ga/gaat/gaan) het met jou?" ("How (are/is/to be) you?").
(3) **Syntax exercises:** For syntax exercises, a limited number of constituents are presented in separate blocks in a randomized order. Some of these blocks can be fixed, such as at the beginning or at the end of the sentence, to elicit specific target structures.

2.2 Interaction walkthrough

In Fig. 1 screenshots of the implemented system are shown. The interaction begins with an agent, whose lips and eyes are animated in synchrony with a recorded utterance. This agent starts the dialogs and after the agent stops talking, the response option(s) are shown in the bottom portion of the screen. As discussed in the previous section, the form of these response options depends on the exercise type. The learner responds by choosing and pronouncing one of these options after clicking the "record" button. In Fig. 1A an example screenshot is shown. When users click the record button, they have to utter the whole response and choose the correct word(s) to complete the sentence. The recording is stopped either automatically after a certain period of silence or by clicking the record button again.

If the system is unable to identify the response as one of the options, the learner is encouraged to try again. If the system recognizes the learner's utterance as correct, the corresponding option is highlighted in green and the dialog continues automatically with the next turn, which begins again with the agent speaking. The background photo and ambient background sounds change each time the location of the story changes.

When the system detects one or more errors in the response, the dialog is stopped and the errors are highlighted in red, as shown in Fig. 1B. In this screenshot, a pronunciation exercise is shown in which the system detects an error associated with the grapheme 'eu' (the phoneme

/øː/) in the word ‘nerveus’ (‘nervous’). In the bottom right corner of the screen, three buttons are now active with which users can (1) listen to their attempt, (2) listen to an example of the correct response or (3) continue with the dialog. The user can also click on the blocks highlighted in red to get more information on the error. In the case of pronunciation errors, a recording of an example of the correct sound is played back (both in isolation and within a word). In the case of morphology and syntax exercises, a pop-up window is shown containing textual information on the type of linguistic structure. An example of this ‘language help’ with information about personal and possessive pronouns is shown in Fig. 1C.

At any time, the learner can access a screen containing a ‘scoreboard’ that shows the scores for each of the linguistic target structures via a menu (not shown). This scoreboard is also shown at the end of the dialog. In Fig. 1D, an example of such a scoreboard for the pronunciation exercises is shown. In this case, all of the target sounds were correctly pronounced (or no errors had been made up to that point), except for the ‘eu’ (/øː/). The user can click on each of these labels to go to the appropriate remedial exercises. These remedial exercises are essentially the same as the exercises in the dialog, but they are not presented in a conversational context.

3. Method

3.1 Aim of the evaluation

The prototype system that was developed within the DISCO project was intended to investigate whether speech technology could in principle be employed to enhance L2 learning. It was not intended as a market-ready product; therefore, among other things, the content within the current system is rather limited. However, in the remainder of this paper we will also evaluate these aspects of the system in order to provide directions for improving this specific system and also other ASR-based CALL systems.

The evaluation reported on in this paper was conducted from three different perspectives:

- **A usability review** based on a set of guidelines and heuristics. A usability review will help us to identify certain obvious system flaws.
- **An expert review** based on interviews with domain experts. With the help of an expert review one is more likely to find higher-level issues with the system and its design.
- **A user test** based on teacher and student questionnaires. With the help of a user test we are able to predict the actual problems that might arise during actual use of the system, as well as prioritize problems that were hypothesized in the other evaluations.

For the expert review we chose to use a group of teachers as domain experts. This was in line with the evaluation envisaged in the project proposal and considered relevant by the reviewers. For the user test we selected the same teachers, as well as a group of DL2 learners. The rationale behind this approach is that we regarded the teachers both as domain experts and potential users.

3.2 Usability Review

A usability review is an evaluation of a user interface based on common usability heuristics and best practices. A common set of heuristics is the one presented in Nielsen (1993). A summary of these heuristics is shown in Table A.1. These high-level heuristics are instantiated in more concrete guidelines in Pierotti (1994). The first author performed the usability review by testing the system against the relevant items in Pierotti (1994). He observed 5 DL2 learners who worked with the current version of the system and 10 DL2 learners who worked with previous versions of the system. The focus of this review lies on student-system (rather than teacher-system) interaction.

3.3 Expert Review

This expert review was carried out in the form of semi-structured interviews with independent domain experts. The goal of these expert reviews was to obtain detailed feedback and suggestions. In Section 3.3.1 we describe the participants involved in this study and in Section 3.3.2 we explain how the expert review was performed.

3.3.1 Experts

Nine experts participated in this study. All of them had several years of teaching experience and most of them taught both low-educated and high-educated learners. The experts were affiliated with three different institutes: two regional education and training centres (six experts) and a university language centre (three experts). A regional education and training centre is a combination of institutions from all the sectors of education for adults and senior secondary vocational education.

Several experts were also responsible for the organization of the DL2 department within their institute. This included evaluating and selecting the teaching methods to be used in the courses. Two experts notably had several years of experience developing a DL2 teaching method that is widely used in the Netherlands. None of the experts had any previous experience with ASR-based CALL systems.

3.3.2 Procedure

For the purpose of this research, the experts participated in a session comprising (1) an introduction to the DISCO system, (2) a questionnaire and (3) and an interview. We had three individual sessions, one session with two experts and one session with four experts. Before the session, the experts were sent a document in which the purpose of the DISCO system was explained, as well as a short description of the exercises within the system.

In the beginning of the session, which lasted 90 minutes on average, the system was introduced to the experts, together with supporting movie clips of users working with the system.

After this introduction, the experts were able to work with the system by themselves. At all times the experts had the opportunity to ask questions. Then the experts completed a questionnaire (see Section 3.4.1). Afterwards, this questionnaire was used as the basis for an open-ended interview in which the researcher asked the experts to explain their answers on the questionnaire.

3.4 User testing

We designed questionnaires to evaluate the system from a user's perspective. Both teachers and students are considered users of the system, albeit from different perspectives. For this reason, we investigated how both DL2 teachers and students experienced the system.

3.4.1 Teacher testing

The Unified Theory of Acceptance and Use of Technology (UTAUT) model, presented in Venkatesh et al. (2003), is a more recent version of their initial Technology Acceptance Model. The aim of the UTAUT model, like that of other technology acceptance models, is to predict the user's intention to use an information system and subsequent usage behavior. The model states that four key constructs: (1) performance expectancy, (2) effort expectancy, (3) social influence and (4) facilitating conditions are direct determinants of usage intention and behavior. Furthermore, the model states that the gender, age, experience and voluntariness of use mediate the impact of the four key constructs on usage intention and behavior.

Because of its relative success in predicting real usage behavior (Kijsanayotin, Pannarunothai & Speedie, 2009; Im, Hong & Kang, 2011), we have used the UTAUT model

in our research to develop the teacher questionnaire. For usage intention and the four direct determinants assumed in the UTAUT model we developed a number of questions. During the sessions described in Section 3.3, the teachers were instructed to indicate the extent to which they agreed with these items using a 7-point Likert scale. The questionnaire is shown in Table A.2.

3.4.2 Student testing

For the purpose of finding problems that actual language learners might have with the system, we requested five DL2 students at the CEFR A2 level to work with the system. The students were all high-educated females, their ages ranged from 18 to 36 and their L1s are English (2x), Chinese, Farsi and Armenian.

After they read a short manual, which was available in both English and Dutch, the students worked with the system individually for 45 minutes. A researcher was present to observe the students interacting with the system. The structure of the session is shown in Table A.3. Afterwards, the students filled in a questionnaire (shown in Table A.4), and were encouraged to give suggestions to improve the system.

4. Results

4.1 Usability and Expert Review

The results of the usability review are presented in Table A.2 in the Appendix. The comments in this table are categorized according to the heuristic in Table A.1 that they pertain to. Not every heuristic in Table A.1 is included because for some of these there are no relevant comments.

The results of the Expert review have been structured using six subsections: (1) pronunciation exercises, (2) morphology and syntax exercises, (3) user interface, (4) content, (5) low-educated learners and (6) practical considerations, which are presented below.

4.1.1. Pronunciation exercises

All experts agreed that students can learn Dutch pronunciation better with the system than without it. This is mainly based on the fact that currently they cannot spend a lot of time on pronunciation within their lessons although they think that it is important. Furthermore, they do not know of any real possibilities for students for practicing pronunciation at home. Two applications are mentioned with which students can practice pronunciation by repeating and replaying their own utterances, but the experts think these programs are less valuable than the DISCO system because they do not give the feedback required.

One expert experienced problems in the feedback for some specific vowel sounds. This is probably due to the regional variety of Dutch spoken by the expert. Although the pronunciation error detection algorithms are trained using data from a large number of speakers who speak different varieties of Dutch, there is a limit to the amount of variation that the system considers correct.

Some experts think that the corrective feedback the students receive might not be sufficient to solve their pronunciation problems. Two experts specifically argue that once a problem is found in the context of the dialog, this problem should afterwards be dealt with in isolation before it can be brought back into the context. Since there can be different causes for a “pronunciation” problem such as (1) an erroneous grapheme-to-phoneme mapping, (2) difficulty in the auditory discrimination of certain sounds and (3) a production problem, the experts suggest adopting different strategies for the various errors. For instance, if the cause of the problem can be found, an appropriate piece of information or exercise should be

offered. This could for example be a sound discrimination exercise, a video/animation showing how to pronounce certain sounds or a production exercise.

4.1.2. Morphology and syntax exercises

Some experts estimate that the added value of the morphology and syntax exercises is lower than that of the pronunciation exercises. These experts state that they already pay a lot of attention to these topics in their lessons using textual exercises. On the other hand, all experts think that exercises that make use of the spoken modality are different from textual exercises; i.e., they think that students will learn something extra or different from the exercises using spoken output. One expert thinks that the exercises that make use of predefined 'blocks' of text are useful from a pedagogical perspective in the sense that in this way problematic constructions can be elicited and tested in a structured manner.

The four types of error classes in both the morphology and syntax exercises are considered adequate by all experts. One expert suggests that exercises related to the conjugation of past tense verbs be added.

4.1.3. User interface

All experts agree that the system is relatively easy to use and they were all able to work with the program after a short introduction. One expert says that a video tutorial or on-screen instructions in the first session would be useful for most students. Some experts experienced problems with the automatic end-of-sentence detection, which sometimes stopped the recording before they were finished speaking.

One expert proposes that, in case of detected learner errors, only the word with the error should be repeated, which would be less annoying than repeating the whole sentence. This technique would also isolate the student's problem. One expert mentions that the

experience after successfully completing a dialog turn or remedial exercise could be made more rewarding in order to increase motivation. At this moment, the system reacts by coloring the prompt green and by automatically proceeding to the next turn or exercise. The expert suggests that a score bar representing the overall current performance of the student could be shown. Another expert argues that the language help, which can only be accessed after an error has been made, should be accessible at all times. Furthermore, the linguistic information should be formatted more clearly. One expert indicates that he thinks the automatic lip synchronization contains some errors and that this can be disturbing for the student.

4.1.4. Content

We define the content of the system as the collection of all dialogs, remedial exercises and language help. The situations and topics in the dialogs and remedial exercises are considered suitable for the target student population. However, the language help, accessible after an error has been made, contains terminology that is possibly unknown even to high-educated learners. The experts suggest that the language help be based mainly on examples of correct and incorrect examples of language use.

Several experts indicate that in some cases the response options in the exercises are too long. This might intimidate some students, forcing them to automatically choose the shorter option if one is available. Furthermore, it would be frustrating to repeat such a response entirely when an error is made.

In the evaluation of the quantity of the content we assumed that the system would be used in parallel with a course of three months with two lessons per week. Most experts think that the number of remedial exercises is large enough to be used in such a course, although some experts would favor more remedial exercises. However, the number of dialogs is considered too small. Most experts suggest that when the system is used during the course one

dialog a week would suffice, resulting in 10-14 dialogs. The dialogs should connect to the themes and linguistic structures that are discussed during the lessons in the course. The experts indicate that most of these themes are comparable across different teaching methods and constitute everyday situations like travelling with public transport, going to the supermarket, going to the bank, hobbies etc. Within these themes different language functions such as agreeing/disagreeing, complaining, greeting, giving opinions, making appointments, invitations, offers, requests, suggestions etc., should be learned. Ideally, all of these should be implemented in the dialogs and remedial exercises.

At the moment, creating and modifying content in the system is not straightforward. The option of creating one's own content is seen as a welcome option by most experts if this were not a time consuming process. One expert suggests that she would like to make her own short dialogs so that her students could practice these at home and that she could discuss them during the following lesson.

4.1.5. Low-educated learners

We asked the experts whether the system would also be suitable for other student populations than the initial target population (see Section 2.1), such as low-educated students. The experts argue that these low-educated learners would need an adapted version of the content. For these students, the difficulty level of the vocabulary used in the dialogs and exercises was considered to be unsuitable. Also the language help was considered unsuitable for them because they are not familiar with the linguistic concepts presented. Furthermore, for these students the dialog interaction should be tightly scaffolded by using, for example, an introductory video so that they know what to do in the dialog.

4.1.6 Practical considerations

Almost all experts have the means to let students use the system, which requires computers with an internet connection and headsets. However, it should be noted that the DISCO system is mainly intended to be used by students at home with minimal help from their teachers. The experts think that this is possible and that both high- and low-educated learners possess the basic computer skills to operate the system. Furthermore, a couple of experts mention that almost all of their students know how to work with a headset because of their familiarity with voice chat software used to communicate with friends and relatives abroad.

Some experts had had negative experiences with using technology in their courses, mostly because of software errors, although the use of technology is encouraged in their work environment. The problem for these experts was that they could not easily solve these problems themselves and that the errors took up valuable class time.

One expert was concerned that she would not easily be able to check whether the system is working properly and that her students might get incorrect feedback. The other experts mostly indicate that they do not feel the need to control all the learning tools their students use and that these tools would probably enhance learning when students practice more by using them.

Most experts expect the students to enjoy working with an interactive system like the DISCO system, which will enhance their motivation. A couple of experts mention that during speaking lessons in the classroom the extroverted students are usually more active and predominant. On the other hand, the introverted students, whose Dutch speaking is usually worse, are less active. These experts argue that for the introverted students, a program like the DISCO system would be especially helpful because they can practice their speaking in a socially safe environment and in an interactive manner.

4.2 User testing

4.2.1 Teacher testing

Fig. 2 shows the histograms of the teachers' answers to the questionnaire. The labels above the histograms coincide with the labels in Table A.3. The first three questions (*PE1*, *PE2*, *PE3*) regarding performance expectancy are all answered positively. That is, the teachers agree that students would better be able to learn the pronunciation, morphology and syntax of Dutch with the DISCO system than without it. The content of the system (*PE4*) as it is right now, is not unanimously deemed sufficient. Regarding the time that could be saved in the classroom by using the system (*PE5*), there seem to be two separate groups of teachers. This can be explained by the fact that teachers who did not think that they would save time generally did not pay a lot of attention to oral proficiency during their lessons. On the other hand, the teachers who spent time on speaking thought that they could gain extra time by using the system. Most of the teachers were not afraid that they would teach their students incorrect information by employing the system in their courses (*PE6*).

The teachers generally think that the system is easy to use (*EE1*), compatible with their current teaching methods (*EE2*) and not frustrating to use (*EE4*). Some teachers do not think that their students have all the knowledge to use the system successfully (*EE3*). This can be ascribed to the fact that the vocabulary is sometimes deemed too difficult and that the learner might not be familiar with the terminology used in the language help (see Section 4.1.4).

Most teachers have the means to use the program (*FC*). The ones who did not state that this is because of the fact that they occasionally work at locations where computers and headsets are not available. The teachers generally agree that, if their budget allowed it, people

in their work environment would support them in using the system (*SC*). All of the teachers would like to use the program (*IOU*).

4.2.2 Student testing

The results of the student testing questionnaire are summarized in Table A.5. Because of the small sample size, we only report the mean answer score for each item. The trend seems to be that the students agree with the positive statements about the system. In summary, the students think that the graphical interface is responsive and visually attractive and that the different types of exercises are helpful and enjoyable. The students rate the system with a 7.8 out of 10.

Besides the questionnaire, the students were also able to give suggestions and other comments. One student mentioned that she found it annoying that she had to repeat the entire sentence when she has made one single error and that she would like to have the ability to go back in the dialog to review her mistake(s). Another student mentioned that it was sometimes irritating that the system "would not hear" her. This referred to the automatic end-of-sentence detection, which sometimes stopped the recording too early or too late. One student recommended adding exercises about the perfect and imperfect tense.

5. Discussion and perspectives for future research and development

We will now discuss the most important findings of this study encompassing both specific feedback about the system as well as more general suggestions on deploying ASR technology in CALL. In general we found that both domain experts and potential users were positive about the performance and user-friendliness of the system. This indicates that the system is currently in such a state that it can empirically be tested on a larger scale. However, we think

that we can learn several important lessons from the three different evaluations. These could be taken into account when improving the present system and when developing other ASR-based CALL systems (see Section 5.1). In addition, there are also remarks that can inspire future research and development (see Section 5.2).

5.1 Discussion

Regarding the validation of the design features, we should first remind the reader that this study focused on a three-pronged evaluation of a prototype resulting from a specified CALL application. The following discussion only applies to this specification and prototyping phase, and not to the earlier mentioned conceptualization stage. As no real-world implementation has happened yet, a validation of the theory behind the concept was not yet possible.

The results of the evaluations indicate that both experts and students have a clear preference for having the students correct problematic elements in isolation rather than having to repeat the whole sentence in which the error appeared. Furthermore, from the usability review we found that after an error has been made and feedback has been provided, it is not always immediately clear which actions could or should be taken by the learner. To help alleviate this potential problem, on-screen pointers can be given during first-time use, which should explain when and why to use a certain functionality.

The experts found the morphology and syntax exercises useful although they already teach these topics using text-based exercises. Apparently the experts think that spoken interaction has an added value compared to written interaction. They also think that by doing these exercises, students are stimulated to speak more in the target language and hereby become more comfortable speaking it. This is in line with arguments adduced to support the output hypothesis in the field of second language acquisition (Swain, 1985; DeBot, 1996),

with views on the importance of speaking practice for improving L2 pronunciation (Kendrick, 1997), and the importance of skill-specific practice for language learning in general (DeKeyser & Sokalski, 1996; DeKeyser, 2007).

Although in the context of DISCO the content was clearly a means rather than a goal in itself, we thought it would be informative to ask questions about the content. The expert review does indeed reveal that the content is a crucial factor in the deployment of CALL systems. This was apparent from the varying opinions of the experts on, for example, the dialog topics, the required vocabulary and the language usage in the language help. These different opinions are caused mainly by differing teacher preferences and students' needs. From this perspective it is clear that there can be no one-size-fits-all CALL system that would help students improve their oral proficiency. It should therefore be possible to both modify existing content, as well as to create new content in a simple manner, using an intuitive interface. However, in relation to ASR this would require specific additional functionalities. The new content should be automatically evaluated in order to assess whether it is appropriate from an ASR point of view. By way of illustration, the words used in an exercise should be available in the lexicon employed by the ASR. Furthermore, these words should not be confusable from an acoustic/phonetic point of view because this makes it difficult for the ASR to keep them apart. Along these lines, the system could discard an exercise in which the words 'ga' and 'gaan' have to be discriminated, but accept an exercise with the words 'jij' and 'jou'. This validation might be implemented by calculating phonetic distances between the possible words or sentences within the exercise and discarding exercises that contain words or sentences that are too similar phonetically. The challenge for the content creator then is to devise exercises that fit the pedagogical needs of the student and are also appropriate from the perspective of ASR.

During the expert review (see Section 4.1.1), some experts stated that the corrective feedback the students receive might not be sufficient to solve their pronunciation problems and that the learning process could be more structured around a pedagogical strategy that adopts knowledge about what causes problems in oral proficiency (see also, Engvall & Bälter, 2007). Different problems should be addressed by adopting different strategies and by offering different remedial exercises. The current DISCO system was not designed with these different types of pedagogical strategies in mind, but in principle it is possible to add different training strategies depending on the nature of the error. This could be achieved in a very simple, deterministic way by deciding beforehand which errors belong to which category and by relating error category to training strategy and remedial exercises. Alternatively, in a more advanced system error categorization could be one of the tasks performed by the system itself, but it is clear that this would require further research, as will be explained in the following section.

5.2 Future research and development

In a more intelligent ASR-based CALL system it should be possible to include tailored, diagnostic exercises and tests to establish which of the three possible sources mentioned in the previous section caused a pronunciation error. The system would then assess performance in such exercises and connect the results to possible training strategies and remedial exercises. While for certain pronunciation errors it might be relatively easy to establish the cause, for others this may be highly complex, which implies that the development of appropriate diagnostic tests would require further research.

Further research would also be necessary to improve the existing DISCO system with respect to a number of points. As explained in the introduction, the objective of the evaluations reported on in this paper was not to assess the performance of the ASR and pronunciation error detection modules in isolation. Rather, we evaluated the system as a

whole. The participants who worked with the system were generally positive about the ASR and error detection performance. However, there are still problems that need to be investigated. First, as discussed above, in the DISCO system ASR performance is heavily dependent on the content of the exercises. There are still exercises in the current system that should be modified or probably be removed to achieve better performance.

Second, one expert noted that, within the pronunciation exercises, the system sometimes detected errors in her utterances, even if she thought she did not produce incorrect sounds. In the expert review we mentioned that this possibly was caused by the regional variety of Dutch that this expert speaks. This touches upon one of the most central problems in automatic pronunciation error detection, namely defining what should be considered 'correct' and what should be considered 'incorrect' pronunciation.

Most experts agreed which target sounds are problematic for DL2 learners, and they thought that these are mostly in line with those addressed in the DISCO system. However, human listeners do not always agree about which of these target sounds contain pronunciation errors in non-native speech (van Doremalen et al., 2012), which causes problems in the development and evaluation of pronunciation error detection algorithms. Moreover, we found that most DL2 teachers do not have clear-cut ideas about when and how to give feedback on pronunciation problems. It is therefore very difficult to evaluate how pronunciation error detection algorithms perform without the context of a real application, and it is not clear what the impact of the technical performance of error detection algorithms is on the learning process. Note that the algorithms used for pronunciation error detection are also relevant for the detection of morphological errors, as in Dutch the latter often manifest themselves as slight acoustic variations of the target form (for example, the presence of absence of schwa, /t/ and /n/). We envisage that by testing the current DISCO system on a larger scale and by monitoring the system usage and user feedback, the most important and tenacious problems

with error detection and ASR become apparent and that these problems can in turn be addressed in the context in which solving them can directly improve the learning process.

6. Conclusions

In this research we have evaluated the DISCO ASR-based CALL system from three different perspectives. From these evaluations, we can conclude that domain experts and users (DL2 teachers and students) are generally positive about the system and intend to use it if they get the opportunity. Several recommendations have been made to improve the system, which range from specific changes and additions to the system to more general statements about the pedagogical and technological issues involved.

Important conclusions are that spoken interaction is considered to have an added value compared to written interaction, that there is no one-size-fits-all CALL system to help students improve their oral proficiency and that it would be important to add different training strategies depending on the nature of the errors to be addressed.

These recommendations can be used to improve the DISCO system and to develop other ASR-based CALL systems so that they can be deployed and tested in a real-life setting, as well as be used as a research tool to investigate language learning processes.

7. Acknowledgements

The DISCO project was funded by the Dutch and Flemish Governments through the STEVIN programme (http://taalunieversum.org/taal/technologie/stevin/). We would like to thank the experts from Radboud in'to Languages, Arcus College and ROC Nijmegen for their valuable feedback and the students who participated in this research for their cooperation. We are indebted to two anonymous reviewers for their useful comments.

8. References

Benzeghiba, M. De Mori, R., and Deroo, O. (2007). Automatic speech recognition and speech variability: a review. *Speech Communication*, 49, 763–786.

Bodnar, S.E., Cucchiarini, C., & Strik, H. (2011). Computer-assisted grammar practice for oral communication. *Proceedings of the Third International Conference on Computer Supported Education*, Noordwijkerhout, The Netherlands, 355–361.

Chevalier, S. (2007). Speech interaction with Saybot, a CALL software to help Chinese learners of English. *Proceedings of the SLaTE-2007 workshop*, 37-40.

Colpaert, J. (2010). Elicitation of language learners' personal goals as design concepts. *Innovation in Language Learning and Teaching*, 4(3), 259-274.

Colpaert, J. (2013). The role and shape of speech technologies in well-designed language learning environments. *Proceedings of the SlaTE-2013 workshop*, 16-19.

Colpaert, J. (2014). Educational Engineering and Distributed Design. Research Report. www.jozefcolpaert.net/EE.pdf

Cucchiarini, C., van Doremalen, J. and Strik, H. (2012). Practice and feedback in L2 speaking: an evaluation of the DISCO CALL system. *Proceedings of Interspeech 2012.*

De Bot, K. (1996). The psycholinguistics of the Output Hypothesis. *Language Learning*, 46(4), 529-555.

DeKeyser, R.M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, 46(4), 613–642.

DeKeyser, R. (2007). Practice in a second language, *Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation, *System*, 36, 506–516

Dörnyei, Z., & E. Ushioda (2009). *Motivation, language identity and the L2 self*. Bristol: Multilingual Matters.

van Doremalen, J., Cucchiarini, C., & Strik, H. (2010). Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing.*

van Doremalen, J., Cucchiarini, C., & Strik, H., (2011). Automatic Speech Recognition in CALL systems: The essential role of adaptation. *Communications in Computer and Information Science*, 126, 56-69.

van Doremalen, J., Cucchiarini, C., & Strik, H. (2013). Automatic pronunciation error detection in non-native speech. *Journal of the Acoustical Society of America,* 134(2), 1336-1347.

van Doremalen, J., (2014) Developing Automatic Speech Recognition-enabled language learning applications: from theory to practice. PhD Thesis, Radboud University Nijmegen.

Ellis N.C. & Larsen-Freeman, D. (2009). Constructing a Second Language: Analyses and Computational Simulations of the Emergence of Linguistic Constructions From Usage. *Language Learning,* 59(1), 90-125.

Engwall, O., & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers, *Computer Assisted Language Learning*, 20(3), 235-262.

Eskenazi, M. (1996). Using Automatic Speech Processing for Foreign Language Pronunciaton Tutoring: Some Issues and a Prototype. *Language Learning & Technology*, 2(2), 62-76.

Eskenazi, M., Kennedy, A., Ketchum, C., Olszewski, R., & Pelton, G. (2007). Carnegie Speech NativeAccent The NativeAccentTM Pronunciation Tutor: Measuring Success in the Real World. *Proceedings of the SLaTE-2007 workshop.*

Franco, H., Bratt, H., Rossier, R., Venkata, R.G., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3), 401-418

Golonka, E.M., Bowles, A.R., Frank, V.M., Richardson, D.L., & Freynik, S. (2012). Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer Assisted Language Learning*, 27(1), 70-105.

Hubbard, P. (2002). Interactive Participatory Dramas for Language Learning. *Simulation and Gaming*, 33, 210-216.

Im, I., Hong, S. and Kang, M.S. (2011). An international comparison of technology adoption: testing the UTAUT model. *Information & Management,* 48(1), 1-8.

Johnson, W.L., Marsella, S., Mote, N., Viljhjalmsson, H., Narayanan, S. & Choi, S. (2004). Tactical language training system: supporting the rapid acquisition of foreign language and cultural skills. *Proceedings of ICALL-2004.*

Kendrick, H. (1997). Keep them talking! A project for improving students' L2 pronunciation, *System*, 25(4), 545-560.

Kijsanayotin, B., Pannarunothai S., Speedie, S.M. (2009). Factors influencing health information technology adoption in Thailand's community health centers: Applying the UTAUT model. *International Journal of Medical Informatics*, 78(6), 404-416

Morton, H. & Jack, M.A. (2005). Scenario-Based Spoken Interaction with Virtual Agents. *Computer Assisted Language Learning*, 18(3), 171-191.

Morton, H. & Jack, M.A. (2010). Speech interactive computer-assisted language learning: a cross-cultural evaluation, *Computer Assisted Language Learning*, 23(4), 295-319.

Neri, A., Cucchiarini, C. and Strik, H. (2006). Selecting segmental errors in L2 Dutch for optimal pronunciation training. *International Review of Applied Linguistics*, 44, 357–404.

Nielsen, J. (1994). *Usability Engineering*. San Diego: Academic Press, 115–148.

Pierotti, D. (1994). Heuristic Evaluation - A System Checklist, Xerox Corporation.

Language students and their technologies: Charting the evolution 2006–2011. *ReCALL*, Available on CJO 2013 doi:10.1017/S0958344013000128

Swain, M. (1985). Communicative competence: some roles of comprehensible input and comprehensible output in its development, in Gass, M.A., Madden, C.G. (eds.) *Input in Second Language Acquisition*, Rowley MA: Newbury House, 235-253

Strik, H., Cornillie, F., Colpaert, J., van Doremalen, J., & Cucchiarini, C. (2009). Developing a CALL System for Practicing Oral Proficiency: How to Design for Speech Technology, Pedagogy and Learners. *Proceedings of the SLaTE-2009 workshop.*

Venkatesh, V., Morris, M.G., Davis, G.B., & Davis F.D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27, 425-478.

Witt, S.M. (2012). Automatic Error Detection in Pronunciation Training: Where we are and where we need to go. *Proceedings of IS ADEPT.*

Appendix A

1. Visibility of system status The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
2. Match between system and the real world The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
3. User control and freedom

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
4. Consistency and standards Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
5. Error prevention Even better than good error messages is a careful design which prevents a problem from occurring in the first place.
6. Recognition rather than recall Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
7. Flexibility and efficiency of use Accelerators - unseen by the novice user - may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
8. Aesthetic and minimalist design Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
9. Help users recognise, diagnose, and recover from errors Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
10. Help and documentation Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Table 1. A summary of the usability guidelines presented in Nielsen (1994).

1. Visibility of System Status During the recording of an utterance a waveform is drawn which indicates that the system is recording. Furthermore, the color of the button with which the recording is started is changed. When the recording is stopped and the analysis of the utterance is started the cursor is changed to a spinning clock. When the analysis is completed, usually within two seconds, the cursor is changed back. This sequence of system states is considered to be visualized effectively. When the analyzed utterance is classified as correct, the system briefly shows the correct response in green and proceeds automatically to the following turn or exercise. When the analyzed utterance is classified as incorrect, corrective feedback is given by coloring the relevant text red. In the latter case the dialog stops, but it is not entirely clear what the user should or is able to do next. The red vs. green color distinction that marks correct and incorrect responses might not be sufficiently clear for colorblind users.
2. User Control and Freedom It is not possible for the user to stop or return to the starting screen. Furthermore, within the dialog, it is not possible to go back to the previous dialog turn. These restrictions limit the perceived control of the user over the system. When users are speaking and they already knows that they have made an error, it is not possible to cancel the current utterance. This could lead to some frustration.
3. Consistency and Standards The terminology and icons are consistent throughout the application. The icons used for the various buttons might not be immediately clear.

Therefore, textual button labels might be preferred.
4. Error Prevention A couple of ASR errors were encountered. These errors lead to inappropriate feedback. Some of these types of errors could be avoided by modifying the content of the exercises.
5. Flexibility and Efficiency of Use All possible actions are directly accessible using clickable buttons. Furthermore, the number of possible actions is so small that, once they are known, the system is easy to use for both novice and experienced users. Buttons are provided with mouse-over tooltips
6. Help Users Recognize, Diagnose, and Recover from Errors When ASR errors occur in the syntax and morphology exercises, the users currently are not able to recover from these. When users intend to utter a certain response and it is not recognized as such, they receive inappropriate feedback. An option could possibly be added in which users could override the automatic analysis by manually choosing or dragging the blocks.
7. Help and Documentation Currently, the system contains no documentation that can be accessed from inside the application. Video tutorials and on-screen pointers during first time usage are recommended.

Table 2. Overview of the results of the usability review. The comments are categorized according to the heuristic in Table A.1 that they pertain to. Not every heuristic in Table A.1 is included because for some of these there are no relevant comments.

1. Performance expectancy
PE1. With the program students would better be able to learn the pronunciation of Dutch sounds than

without it. **PE2.** With the program students would better be able to learn Dutch morphology than without it. **PE3.** With the program students would better be able to learn Dutch syntax than without it. **PE4.** I think the dialogs and remediation exercises in the program are comprehensive enough to use the program in my courses. **PE5.** By using the program I could spend more time on other important topics during my lessons. **PE6.** I am afraid students will learn incorrect things when they use the program.
2. Effort expectancy **EE1.** I think the program is easy to use. **EE2.** The program is compatible with the teaching methods I use. **EE3.** Students have the knowledge that is necessary to use the program succesfully. **EE4.** Using the program might be frustrating for me.
3. Facilitating conditions **FC.** I have all the means to use the program (computers with an internet connection, headsets).
4. Social influence **SI.** I think people in my work environment would be helpful I want to use the program.
5. Intention of use **IOU.** If I would get the chance I would use the program.

Table 3. Teacher questionnaire. The items are categorized by the relevant predictors in the UTAUT model.

	Duration	***Activity***
1.	12 minutes	Dialog with syntax exercises
2.	3 minutes	Remedial syntax exercises

3.	12 minutes	Dialog with morphology exercises
4.	3 minutes	Remedial morphology exercises
5.	12 minutes	Dialog with pronunciation exercises
6.	3 minutes	Remedial pronunciation exercises

Table 4. Structure of student testing session.

	Answer range	***Mean***
General		
1. The instructions are clear.	*1 = totally disagree – 4 = totally agree*	3.00
2. I think the buttons on the screen, the mouse and the keyboard are easy to use.	*1 = totally disagree – 4 = totally agree*	3.50
3. I don't like speaking into the microphone.	*1 = totally disagree – 4 = totally agree*	2.00
4. The program is fast enough.	*1 = totally disagree – 4 = totally agree*	3.25
5. I think the program is visually attractive.	*1 = totally disagree – 4 = totally agree*	3.75
6. It helps me that I can replay my recording.	*1 = totally disagree – 4 = totally agree*	3.25
7. It helps me that I can listen to an example	*1 = totally disagree – 4 = totally agree*	3.75
8. I think the dialogs are fun to do.	*1 = totally disagree – 4 = totally agree*	3.75
9. I think the dialogs are realistic.	*1 = totally disagree – 4 = totally agree*	3.50
Morphology exercises		
10. I think the morphology exercises are fun to do.	*1 = totally disagree – 4 = totally agree*	3.25
11. I understand the feedback in the morphology exercises.	*1 = totally disagree – 4 = totally agree*	3.00
12. I learn something from the feedback in the morphology exercises.	*1 = totally disagree – 4 = totally agree*	3.00
13. The extra theoretical information with the morphology exercises is a good help.	*1 = totally disagree – 4 = totally agree*	2.67
14. I learn something from the extra morphology exercises.	*1 = totally disagree – 4 = totally agree*	3.00
15. I think the morphology exercises are:	*1 = too difficult – 5 = too easy*	2.25

Syntax exercises		
16. I think the syntax exercises are fun to do.	*1 = totally disagree – 4 = totally agree*	4.00
17. I understand the feedback in the syntax exercises.	*1 = totally disagree – 4 = totally agree*	3.00
18. I learn something from the feedback in the syntax exercises.	*1 = totally disagree – 4 = totally agree*	3.00
19. The extra theoretical information with the syntax exercises is a good help.	*1 = totally disagree – 4 = totally agree*	3.50
20. I learn something from the extra syntax exercises.	*1 = totally disagree – 4 = totally agree*	3.00
21. I think the syntax exercises are:	*1 = too difficult – 5 = too easy*	2.33
Pronunciation Exercises		
22. I think the pronunciation exercises are fun to do.	*1 = totally disagree – 4 = totally agree*	3.25
23. I understand the feedback in the pronunciation exercises.	*1 = totally disagree – 4 = totally agree*	3.75
24. I learn something from the feedback in the pronunciation exercises.	*1 = totally disagree – 4 = totally agree*	3.50
25. The extra theoretical information with the morphology exercises is a good help.	*1 = totally disagree – 4 = totally agree*	3.25
26. I learn something from the extra pronunciation exercises.	*1 = totally disagree – 4 = totally agree*	3.25
27. I think the pronunciation exercises are:	*1 = too difficult – 5 = too easy*	3.00
Overall appreciation		
28. Would you use the program?	*no = 0, yes = 1*	1.00
29. What grade (from 1 to 10) would you give to the program?	1 - 10	7.75

Table 5. Student questionnaire. The first column contains all the items. The second column shows the answer range per item. In the third column, the means of the scores given by the participants are shown.