Practice and feedback in L2 speaking: an evaluation of the DISCO CALL system

Catia Cucchiarini, Joost van Doremalen, Helmer Strik

Centre for Language and Speech Technology, Radboud University, Nijmegen, The Netherlands

{c.cucchiarini;j.vandoremalen;h.strik}@let.ru.nl

Abstract

In this paper we report on the ASR-based CALL system DISCO: Development and Integration of Speech technology into COurseware for language learning. The DISCO system automatically detects pronunciation and grammar errors in Dutch L2 speaking and generates appropriate, detailed feedback on the errors detected. We briefly introduce DISCO and present the results of a first evaluation of the complete system.

Index Terms: Computer Assisted Language Learning, ASR, speaking proficiency.

1. Introduction

In recent years, the interest in applying Automatic Speech Recognition (ASR) technology to second language (L2) learning has been growing considerably [6] because of the possibility of assessing L2 oral skills and providing corrective feedback automatically. Developing ASR-based systems that can provide accurate and useful feedback on oral proficiency is not trivial though, because L2 learner speech poses special difficulties to ASR technology [1], [8]. In addition, existing systems in general do not contain the features required. The majority of systems for practicing grammar skills do not support spoken interaction [2], while speech interactive systems that address L2 pronunciation (for an overview of commercial systems, see [9]) in general do not provide immediate, detailed feedback on individual segments in connected speech.

To fill this gap a project (DISCO: Development and Integration of Speech technology into COurseware for language learning) was started aimed at developing an ASR-based system that automatically detects pronunciation and grammar errors in Dutch L2 speaking and that generates appropriate, detailed feedback on the errors detected. In the remainder of this paper we first briefly introduce the DISCO system (Section 2), we then present a first evaluation of the complete DISCO system (Sections 3 and 4), discuss the results (Section 5) and draw conclusions (Section 6).

2. The DISCO system

The aim of DISCO was to develop a prototype of an ASR-based CALL application for Dutch L2. The application aims at optimizing L2 learning through interaction in realistic situations and at providing intelligent feedback on important aspects of L2 speaking such as pronunciation, morphology, and syntax. The application is able to detect errors and to provide feedback on the errors made by DL2 learners.

2.1. The design of DISCO

In DISCO, we limited our general design space to closed response conversation simulation courseware and interactive participatory drama, a genre in which learners play an active role in a pre-programmed scenario by interacting with computerized characters or "agents". Information on appropriate feedback strategies, pedagogical goals and personal goals was obtained through focus group discussions and was then taken into account in finalizing the DISCO design [7].

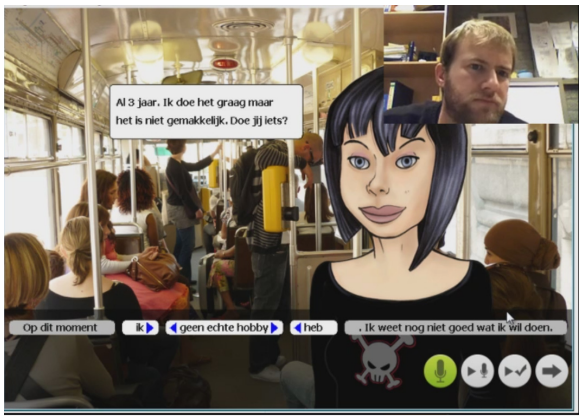

Figure 1. *Example of a syntax exercise, with in the upper-right corner a language learner using the system*

The learning process starts with a simulation of a realistic conversation in which students can choose from a number of prompts at every turn. Based on their errors they are offered remedial exercises, which are very specific and constrained exercises. Feedback depends on individual preferences: the default strategy is immediate corrective feedback visually implemented through highlighting, which puts the conversation on hold and focuses on the errors. Learners who wish to have more conversational freedom can choose to receive communicative recasts as feedback, which let the conversation go on while highlighting mistakes for a short period of time.

2.2. Speech recognition and error detection

Based on the exercises described in the previous section, we designed a system architecture according to the requirements stated during the courseware design phase. To handle the students' utterances a two-step procedure is employed: first it is determined what was said (speech recognition), and second how it was said (error detection). The speech recognition module determines the sequence of words the student uttered. For each prompt a list of predicted correct and incorrect responses is created beforehand based on errors that are expected on empirical grounds. This list is the basis for a Finite State Grammar (FSG) language model, which is used by a Hidden Markov Model (HMM)-based speech recognition system. The recognition system is forced to choose among the predicted responses from the list.

After the selection of the best matching utterance, the utterance verification step is needed to verify whether the selected response was actually uttered by the learner. This is done through a confidence measure based on the acoustic likelihood of the utterance. In this way, syntactical errors and some morphological errors are detected through speech recognition, so that no additional analysis is needed. For pronunciation errors an additional analysis is required.

The canonical phone string (target pronunciation) is encoded in a weighted FSG, together with frequently observed pronunciation errors which are represented in parallel arcs. The arcs carrying pronunciation errors have a certain transition cost assigned to them, in order to keep the number of false alarms acceptable.

3. Evaluation of the DISCO system: Method

A first evaluation of the whole system was conducted to gain insight into aspects such as user satisfaction and feedback accuracy. Groups of DL2 students in Antwerp, Flanders, and Nijmegen, the Netherlands, worked with DISCO and filled in a questionnaire that measured their satisfaction with the system. The student-system interactions (system prompts, student responses, system feedback, etc.) were recorded and experts analyzed them to check the quality of the feedback provided by the system on pronunciation, morphology and syntax.

3.1. Subjects

The DISCO program was evaluated by a total of 23 students (6 males and 17 females), 14 at Linguapolis, the language centre of the University of Antwerp and 9 at Radboud in'to Languages, the language centre of Radboud University Nijmegen (9). Three different subgroups evaluated the three components syntax (7), morphology (8) and pronunciation (8). The age of the students varied between 20 and 40. The highest level of education was mostly university level and, in one case, secondary education. The students had different first languages (Farsi, Armenian, Russian, Portuguese, Italian, Spanish, Arabic, Polish, English) and they could all speak one or more foreign languages, most often English, followed by French. The length of stay in the Netherlands or Flanders varied between 4 months and 13 years. There was also a large difference in the amount of time spent learning Dutch, from 4 months to 7 years, but their proficiency level was at or just above CEFR level A2, although such levels are not necessarily consistent across location or rater.

3.2. Procedure

The DISCO evaluation started with an introduction to explain the purpose of the evaluation and to demonstrate the system. Each student worked for at least 20 minutes with a dialogue chosen by the teacher. Based on the mistakes observed during this first part of the evaluation the student was assigned remedial exercises in DISCO for at least 10 minutes. The teacher and the student chose the exercises together based on the report provided by the system. The teacher suggested that the student first work on the topics with the weakest scores.

During the test the teacher only intervened if the student asked for help or if there were technical problems such as disconnections in the speech recognition system. Unfortunately, these occurred during the Nijmegen tests so that these students did not receive feedback from the system in all cases, which affected their evaluation of the program (see 4.5).

After the test the students filled in a questionnaire asking general personal details (age, gender, educational background, country of origin and first language) as well as specific questions related to time spent learning Dutch, computer use, user-friendliness of the program, the quality of dialogues and feedback, the practice exercises and the extra help provided Finally, the students could indicate whether they would choose to use the program themselves, what mark they would give to the program and they could add extra comments.

3.3. Feedback assessment

To assess the quality of the feedback experts listened to recordings of the evaluation experiment and annotated the errors made by the learners. Their annotations were then compared to the feedback the students received from the system. For logging and data collection purposes the system automatically generates Praat [3] text grids containing word alignments, phone alignments and pronunciation errors for each utterance. Experts in Nijmegen and Antwerp listened to sets of responses using these text grids. For syntax and morphology the teachers transcribed the response as they perceived it and their transcriptions were compared to the feedback by the system. For pronunciation the experts listened to the students' responses in the test and indicated the errors they heard. As is well known, such annotations tend to contain an element of subjectivity [4]. The expert transcriptions were then compared with those generated by the speech recognition module.

4. Evaluation of the DISCO system: Results

4.1. Program assessment

In this part of the questionnaire, the students indicated whether they agreed with the following statements:

1. *The system (buttons, mouse and keyboard) is easy to use.*
2. *It is annoying to have to speak into a microphone.*
3. *The microphone worked well.*
4. *The speed of the program is good.*
5. *The program is visually attractive.*

The students were largely positive about the program.. In general they did not have problems using the buttons, the mouse, the keyboard, or the microphone. One student commented that the mouse did not work properly and two students found it annoying to speak into the microphone. Comments on the program's speed were largely neutral. All the students found that the program looked good; two students commented on the attractiveness of the background and the colours. Other student comments were mainly positive (a very good program, interesting), although one student commented that the program made him/her nervous.

4.2. Assessment of dialogues and feedback

In this part of the questionnaire, the students indicated whether they agreed with the following statements:

6. *The dialogues are fun.*
7. *The dialogues are realistic.*
8. *I understand the feedback.*
9. *I learn about syntax/morphology/pronunciation from the feedback.*

The students were positive about the dialogues: they thought they were fun to do and realistic. They were also satisfied with the feedback; they understood it and thought they learnt from it. One student replied that the dialogues were too difficult and another one observed that the system had problems in processing the answer if this was produced too quickly.

4.3. Assessment of practice exercises

After finishing a dialogue the students were presented with a summary of their mistakes and were given the opportunity to practice the areas that needed improving. In general the students indicated that they learnt from the exercises and thought they were fun to do.

10. *The exercises were (1 too difficult, 5 too easy)*
11. *I learnt something from the exercises.*
12. *The exercises were fun to do.*

For syntax the level of difficulty of these exercises was assessed to be "just right". For morphology, on the other hand, the practice exercises were considered to be too easy. For pronunciation, the difficulty level of the practice exercises varied from easy to difficult. One student commented that a more detailed introduction would have been useful while another student found the response time too short.

4.4. Assessment of extra help

Extra help was provided in different forms. Students could listen to a recording of their own utterance, listen to an example utterance (as they should have said it) and first see the correct utterance on the screen by choosing the correct utterance from a number of alternatives.

13. *It is useful to be able to hear myself again.*
14. *It is useful to be able to listen to an example.*
15. *It is useful to be able to drag the squares on the screen (syntax). It is useful to be able to choose the correct option on the screen (morphology) The 'sound answer keys' are very useful in the pronunciation exercises.*

The students found it useful to listen to their own recording and the example utterance, as well as to click on the correct answer.

4.5. Overall assessment

In this section the students answered the following questions:

16. *Would you use the program yourself?*
17. *What mark, from 1 to 10, would you give the program?*

All students said they would use the program themselves. The average mark assigned to the program varied from 9.0 to 7.2 in Belgium and from 8.5 to 5.0 in the Netherlands. This had to do with the defective connection with the speech processor. As one student commented, 'It's good when it works.' Several of the extra comments referred to problems with the interface and 'bugs'. One student found it annoying to have to click so much. Suggestions were also made: fun to have different levels and to have dialogues with other themes. Another student liked the fact that the sentences were first short then got longer.

4.6. Feedback accuracy

4.6.1. Syntax feedback accuracy

For the annotated material, two evaluation measures were calculated:

1. The percentage of utterances with correct feedback: this indicates the proportion of utterances where the system gave the correct type of feedback, i.e. the utterance contains errors or the utterance does not contain errors.
2. The percentage of utterances correctly recognized: this indicated the proportion of utterances where the system recognized the utterance correctly in terms of the sequence of words. Disfluencies such as false starts and repetitions of words are not taken into account, as well as phonetically similar pronunciation variants ('me' vs 'mij', 'we' vs 'wij').

Table 1. *Evaluation measures for feedback on syntax in Antwerp and Nijmegen*

	Antwerp	*Nijmegen*
Number of annotated utterances	193	74
Number of subjects	5	2
% Utterances with correct feedback	87.6%	82.4%
% Utterances correctly recognized	80.3%	77.0%

In the syntax exercises, blocks (words or groups of words) have to be uttered in the syntactically correct order. When these blocks are too short, different permutations of these blocks can be easily confused by the speech recognizer. This is especially the case when the utterance also contains filled pauses or other disfluencies and speaker sounds which can be misinterpreted as short words such as 'ik' (I) and 'me' (me). This can be solved by changing these problematic exercises.
Other errors were caused by the students starting to talk before pressing the 'start recording' button or pressing the 'stop recording' button before finishing the whole sentence.

4.6.2. Morphology feedback accuracy

Most system errors in the morphology exercises can be ascribed to the failure of the speech recognizer to discriminate between two or more phonetically highly similar morphological variants. This is the case for verbs with or without an ending schwa or /t/, 'we' vs 'wij', 'me' vs 'mij' etc. Technically, these exercises should be avoided because they are too error-prone.

Table 2. *Evaluation measures for feedback on morphology in Antwerp and Nijmegen*

	Antwerp	*Nijmegen*
Number of annotated utterances	207	67
Number of subjects	5	2
% Utterances with correct feedback	80.2%	73.1%
% Utterances correctly recognized	71.5%	73.1%

4.6.3. Pronunciation feedback accuracy

For the feedback provided by the DISCO system on the pronunciation exercises we calculated four measures:

- **Correct Accept**: the number of sounds marked as correct by the system and by the annotator
- **False Accept:** the number of sounds marked as correct by the system but not by the annotator

- **False Reject**: the number of sounds marked as erroneous by the system but not by the annotator
- **Correct Reject**: the number of sounds marked as erroneous by the system and by the annotator

Table 3.*Evaluation measures for feedback on pronunciation in Antwerp and Nijmegen*

	Antwerp, 2 subjects 36 utterances	Nijmegen, 5 subjects 81 utterances
CA	304	711
FA	4	4
FR	14	92
CR	31	27
Precision CA	98.7%	99.4%
Precision CR	68.9%	22.7%
Recall CA	95.6%	88.5%
Recall CR	88.0%	87.1%

These data reveal that there are many more pronunciation errors in Antwerp than in Nijmegen, which is surprising given that the students in the two groups had the same proficiency level. Most errors appear to be false rejects, especially in Nijmegen, which leads to a low value for Precision CR. This point deserves further attention (see below).

5. Discussion

The results of the evaluation presented in the previous sections provide an overall positive picture: in general the students appear to appreciate the system and most of its features, although there is clearly room for improvement. The analyses of feedback accuracy also return generally positive results, but they make it very clear that also in this respect there is room for improvement.

It is obvious that the evaluation of the program as a whole is also related to the performance of the technology, which is apparent from the comments given by the students in Nijmegen, where the connection failed at various points and the system could not provide feedback. This is a shortcoming of the present evaluation which limits its informative power. For these and other reasons we are now planning a new round of evaluations with an improved version of the system in which disconnections will not occur. However, for improving the system and for conducting these future evaluations it is interesting to analyse the results of the present one in more detail to see what we can learn from it.

In the syntax exercises most inaccuracies were caused by short words not being recognized correctly. Similarly, for morphology it appeared that distinctions that hinge on subtle acoustic differences, like the presence or absence of schwa, /t/, and /n/ to distinguish different grammatical forms are problematic. For usability purposes such difficult aspects should be avoided.

An intriguing point about the present results is that the relatively high number of pronunciation false rejects, especially in Nijmegen, did not necessarily lead to negative evaluations on the part of the students. There are various possibilities: the students did not notice that they received "erroneous" feedback, they assumed they were making errors; this feedback was not "erroneous" after all.

To gain insight into the discrepancy in number of errors between Antwerp and Nijmegen, we interviewed the two annotators. It appeared that while the annotator in Antwerp had checked whether the system's feedback was appropriate, the Nijmegen annotator had made her own annotations independently of the system. Although this might be another interesting way of evaluating feedback accuracy, it was not exactly the procedure we intended to adopt, given that our aim was to determine whether the feedback by the system was appropriate or not. Further inspection of the false rejects in Nijmegen revealed that 29 out of 92 concerned the distinction /a:/ - /ɑ/, a difficult distinction to categorize in L2 speech [8]. In other words, it is not exactly clear whether these false rejects are due to real inaccuracies by the system or to a less strict evaluation by the annotator. In any case, for future evaluations it seems that we will have to involve more than one annotator to get less subjective expert assessments. In addition, we intend to optimize the parameters in the current pronunciation error detection setup using the speech material collected in this evaluation.

6. Conclusions

Our first evaluation of the complete DISCO system was generally positive about the system and the technology used, but also provided clear indications on how to improve both. Certain syntax and morphology exercises that rely on subtle differences can better be avoided, while further experiments are required to get better insight into the accuracy of pronunciation feedback, for fine-tuning the technology and obtain better performance.

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 whole DISCO team and in particular Ghislaine Giezenaar (Radboud in'to Languages) and Liesbeth Melis (Linguapolis) for organizing and conducting the evaluation tests.

8. References

[1] Benzeghiba, M., Mori de, R., Deroo, O. Dupont, S., Erbes, T., Jouvet, D., Fissore, L. Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. "Automatic speech recognition and speech variability: a review", Speech Communication 49, 763–786, 2007.

[2] Bodnar, S., Cucchiarini, C. Strik, H. "Computer-assisted grammar practice for oral communication", Proc. of the 3rd International Conference on Computer Supported Education (CSEDU), 2011.

[3] Boersma, P. and Weenink, D. Praat: doing phonetics by computer (Version 5.1.10) [Computer program]. Retrieved July 8, 2009, from http://www.praat.org/.

[4] Cucchiarini, C., Assessing transcription agreement: methodological aspects, Clinical Linguistics & Phonetics, 10 (2), 131-155, 1996.

[5] Demuynck, K., Roelens, J., Van Compernolle, D., and Wambacq, P., "SPRAAK: An Open Source SPeech Recognition and Automatic Annotation Kit". In Proc. Interspeech, 495-498, 2008.

[6] Eskenazi, M., "An overview of Spoken Language Technology for Education", Speech Communication, 2009.

[7] Strik, H., Colpaert, J., van Doremalen, J. & Cucchiarini, C. The DISCO ASR-based CALL system: practicing L2 oral skills and beyond, Proceedings LREC 2012, Istanbul, Turkey, 2012.

[8] Van Doremalen, J., Cucchiarini, C., Strik, H., Automatic pronunciation error detection in non-native speech (submitted).

[9] Witt, S. "Automatic error detection in pronunciation training: Where we are and where we need to go", Proc. IS ADEPT, Stockholm, Sweden, 2012.