ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Developing a CALL System for Practicing Oral Proficiency: How to Design for Speech Technology, Pedagogy and Learners

Helmer Strik [a], Frederik Cornillie [b], Jozef Colpaert [b], Joost van Doremalen [a], Catia Cucchiarini [a]

[a] Department of Linguistics, Radboud University, Nijmegen, The Netherlands
[b] Linguapolis - Institute for Language and Communication, University of Antwerp, Antwerp, Belgium
h.strik|j.vandoremalen|c.cucchiarini@let.ru.nl;frederik.cornillie|jozef.colpaert@ua.ac.be

Abstract

Automatic recognition of non-native speech is problematic. A key challenge in developing spoken CALL systems is to design exercises that enable learning but which are still technically feasible. This especially applies to systems intended for practicing grammar. In the current paper we focus on the issue of matching design and speech technology. On the one hand we are developing and testing speech technology modules to determine what is feasible. On the other we use this knowledge in designing a CALL system for practicing pronunciation and grammar.

1. Introduction

Current theories of second language acquisition emphasize the importance of performance as a means of acquiring competence in a second language: if learners want to speak a second language fluently and accurately, it is necessary for them to practice speaking it [1]. For speaking proficiency achieving a sufficient amount of practice in the classroom can be difficult owing to lack of time. Recent developments in automatic speech recognition (ASR) have opened up opportunities of developing CALL systems for oral skills. Realizing ASR-based CALL systems that can provide training and feedback for L2 speaking is not trivial, as ASR performance on non-native speech is not yet as good as on native speech. To circumvent at least part of the ASR problems caused by non-native speech, various techniques have been proposed. One of them consists in eliciting constrained output from L2 learners. However, this should be done carefully as too many constraints may affect the communicative nature of the L2 learning program. In addition, L2 learners need to have some freedom in formulating answers when practicing grammar in speaking proficiency in order to show whether they are able to produce correct forms. So, the challenge in developing an ASR-based system for practicing oral proficiency consists in designing exercises that allow some freedom to the learners in producing answers, but that are predictable enough to be handled by ASR.

This is precisely the challenge we face in the DISCO project [2], which is aimed at developing a prototype of an ASR-based CALL application for practicing speaking performance in Dutch as a second language (DL2). The application aims at optimizing learning through interaction in realistic communication situations and at providing intelligent feedback on important aspects of DL2 speaking, viz. pronunciation, morphology, and syntax. The application should be able to detect and give feedback on errors that are made by learners of Dutch as a second language

Such an application requires dedicated speech technology modules for non-native speech recognition and error detection and, of course, expertise in different fields. The speech technology modules do not stand on their own, but have to be embedded in the whole system, and their suitability is related to the goals, the targeted users, and the feedback moves.

In this paper we present the results of preparatory studies we conducted to finalize the design and the development of the speech recognition modules. We further discuss their consequences and the importance of taking various factors into account when designing ASR-based CALL applications.

2. Design

In this section we describe the design of the DISCO application and present the results of a number of preparatory studies we carried out to gain more insight into appropriate feedback strategies, pedagogical goals and personal goals.

In DISCO, we limit our general design space to closed response conversation simulation courseware and interactive participatory drama (IPD), a genre in which learners play an active role in a pre-programmed scenario by interacting with computerized characters or “agents”. The use of drama is beneficial for various reasons, a) it “reduces inhibition, increases spontaneity, and enhances motivation, self-esteem and empathy” [3], b) it casts language in a social context and c) its notion implies a form of planning, scenario-writing and fixed roles, which is consistent with the limitations we set for the role of speech technology in DISCO. To summarize, this framework allows us to create a rich and communicative CALL application that stimulates DL2 learners to produce speech and experience the social context of DL2. On the other hand, these choices are safe from a technological perspective, and are appropriate for successfully deploying an ASR while taking into account its limitations.

2.1. Interviews with DL2 teachers and experts

Exploratory in-depth interviews with DL2 teachers and experts were conducted. The results presented in this sub-section concern their opinions about DL2 learners.

Two types of DL2 learners were identified: those who want immediate corrective feedback on mistakes, and those who want to proceed with conversation training even if they make mistakes. Teachers also believed that our target group (highly-educated DL2 learners) will probably prefer immediate corrective feedback. To cater for both types of learners, the system could provide two types of feedback

strategies and have the learners choose the one that suits them better through parameter setting.

DL2 learners often want more opportunities to practice. A CALL system can provide these opportunities. DL2 learners feel uneasy at speaking Dutch because they are not completely familiar with the target language and culture. Therefore, it might be a good idea to provide some information about the target culture(s), so that learners can try to achieve intercultural competence.

2.2.Focus group with DL2 students

Besides the pedagogical goals, the personal goals of learners should be taken into account. A focus group is a qualitative research technique [4] which we used to elicit the personal goals of learners. In this case the focus group consisted of 9 DL2 learners.

DL2 learners often feel discouraged if they don't have sufficient knowledge of the topic of the conversation (politics, habits, etc.). Furthermore, they want to feel respected for their courage to integrate in the target culture(s). The conversations may thus certainly deal with habits and practices of the target culture(s). Also, learners feel frustrated because they cannot keep up with the pace of conversations in the target language.

DL2 teachers and experts mentioned lack of exposure (see 2.1), but the participants did not complain about this lack, even if we explicitly asked them.

2.3.Pilot study with DL2 teachers

The current and the following pilot study were carried out by means of partial systems with limited functionality (e.g. no speech technology). The functions of the system that were not implemented (play prompts, give feedback, etc.) were simulated. For this pilot study, an internet application was used to present one conversation tree (including graphics).

In general, DL2 teachers were positive about the possibilities offered by such a CALL system to practice pronunciation, morphology and syntax. Most of the comments dealt with how the exercises on morphology and syntax should be designed. The main conclusions were that different types of exercises probably require different approaches. For instance, regarding morphology, a multiple choice approach was recommended for personal and possessive pronouns, e.g. "Hoe gaat het met (jij / jou / jouw)?" {How goes it with (you / you / your)?}; but for verb inflections it might be good to present root forms (between brackets), e.g. "Hoe (gaan) het met jou?" {How (to go) it with you?}. For syntax exercises the constituents can be presented in separate blocks, not too many of them (e.g. max. 4), some of these blocks could be fixed and others random (made clear by e.g. using different colors). To test the presence of constituents, e.g. a subject or a pronoun, again another type of blocks (+ color) might be used that are empty or contain optional or multiple (choice) answers.

2.4.Pilot study with DL2 students

A web-based prototype of the application was developed (see http://disco.linguapolis.be/pilot). A teacher simulated the functions that were not yet implemented, e.g. by reading lines from the screen and providing feedback. Audio and video recordings were made and analyzed. The pilot was carried out in Antwerp (5 participants) and Nijmegen (4 participants). The first research question concerned the feedback students prefer. 5 out of 9 respondents indicated a preference for immediate feedback, and 4 out of 9 students responded that they did not know which feedback they preferred. The fact that no student wanted (delayed) feedback confirms the hypothesis that highly-educated learners want to receive overt feedback with high frequency.

Figure 1. Screen shot of the web-based prototype.

In exercises on morphology and syntax students first have to construct the grammatical form they want to utter. As a result, the cognitive load produced by these exercises is probably higher, which in turn may lead to a higher number of disfluencies and to speech recognition and error detection problems. A possible solution might be to ask students to first construct their answer on the screen by means of keyboard and mouse (called textual interaction), and then utter these answers. The average number of disfluencies per turn were calculated for the following 5 cases:

1. pronunciation (no textual interaction);
2. morphology, no textual interaction;
3. morphology with textual interaction;
4. syntax, no textual interaction;
5. syntax with textual interaction.

Table 1. Average values and standard deviations of the disfluency ratios for the 5 cases (see above); no: no textual interaction, with: with textual interaction.

cases	1. no	2. no	3. with	4. no	5. with
Avg.	0.64	0.82	0.34	0.91	0.43
S.D.	0.79	0.45	0.36	0.37	0.15

The average number of disfluencies is significantly smaller in the cases with textual interactions. These results clearly show that this procedure is useful to substantially reduce the number of disfluencies. However, CALL research does suggest that it is beneficial to maintain modalities, and not to use keyboard and mouse interaction in courseware that is essentially conversational in nature [3]. Furthermore, for some students it may not be necessary, or students may have a preference for not using it. Therefore, textual interaction will be optional. If used, we will try to use the output of textual interaction to improve speech recognition and error detection.

Another important result from this pilot study is that the order of events was not always clear to students. Although the

teacher that guided the experiment provided instructions that would normally be shown by the computer, students did things in the wrong order, acted ahead of time, spoke while carrying out the textual interaction, only uttered part of the prompts, or proceeded to the next item without speaking the utterance. The consequences for the design are that we need to clearly structure and scaffold the interaction sequences, give clear and concise instructions, use a push-to-talk button, and only allow students to proceed to the next item if they have finished their task.

Finally, we also noticed that teachers, both in Nijmegen and in Antwerp, spontaneously provided non-verbal feedback during the conversation, and that students clearly responded to this kind of feedback. As CALL research also suggests [5], non-verbal feedback may be used complementarily to the verbal (overt or covert) feedback, and may be beneficial to student motivation and the learning effect. The virtual agents can provide this kind of feedback, e.g. by nodding or shaking their heads, smiling, frowning, etc..

2.5.Consequences for design

The results of the preparatory studies were taken into account in finalizing the design of the CALL system. The learning process starts with a relatively free conversation simulation, taking well into account what is (not) possible with speech technology: learners are given the opportunity to choose from a number of prompts at every turn (branching, decision tree). Based on the errors they make in this conversation they will be offered remedial exercises, which are very specific exercises with little freedom.

Feedback depends on individual learning preferences: the default feedback strategy is immediate corrective feedback, which is visually implemented through highlighting, and from an interaction perspective by putting the conversation on hold and focusing on the mistakes. Learners that 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.

The final system will have several parameters that can be changed by the learner or teacher. During development and implementation, we will try to have these parameters behave intelligently (based on error analysis and learner behavior), so that the system can adapt itself to the learner.

3. Speech technology

In a CALL application, for each prompt the utterances spoken by the DL2 students have to be handled by means of speech technology. In DISCO we intend to adopt a two-step procedure in which

(1) it is first determined what was said (content, speech recognition), and
(2) subsequently how it was said (form, error detection).

In the first phase, which is necessary to establish whether the learner produced an appropriate answer, the system should tolerate deviations in the way utterances are spoken. If the incoming utterance has been identified as being an attempt at producing the required answer, the system proceeds to error detection, the second phase, in which strictness is required (see also [6] and [7]). If the utterance cannot be recognized the system will prompt the user to try again. In the first phase of the two-step procedure two stages can be distinguished: (1a) utterance selection and (1b) utterance verification. We are now developing and optimizing algorithms for these different tasks (see. section 3.3).

3.1.Speech recognition

The system will try to elicit constrained responses by presenting several prompts at each step in the conversation from which the learner can choose one. For each of these prompts (utterances), there will be three versions, for practicing pronunciation, morphology, and syntax. For each version of all prompts there will be a specific list of predicted, correct and incorrect, responses. For instance, for the syntax version the predicted list will contain syntactically correct responses, but also (frequently made) syntactic errors.

The task in the speech recognition phase is to determine which utterance was spoken. In order to do so, a language model is based on the predicted list, and during decoding the optimal path in this language model is chosen. In our experiments we obtained significant improvements by optimizing the language model and the acoustic models and achieved utterance error rates about 8-10%. for stage (1a) utterance selection.

Regarding disfluencies, we found out that filled pauses can be handled well by including 'filled pause'-loops in the language model. Filled pauses are common in everyday spontaneous speech and generally do not hamper communication. The students could therefore be allowed to make (a limited number of) filled pauses. Other disfluencies (restarts, repairs, etc.) are probably more problematic.

In stage (1a), utterance selection, the path in the language model is chosen that best matches the acoustic signal. However, the selected utterance does not always correspond (exactly) to what was actually spoken: the spoken utterance might not be present in the predicted list, or even if it is present it might not end up on position 1. Since giving feedback on the wrong utterance is confusing, we should try to avoid this as much as possible. To this end, confidence measures are calculated in stage (1b) utterance verification.

These confidence measures are compared to optimized thresholds, in order to determine whether the utterance will be accepted or rejected. When the utterance is accepted the learner gets feedback on the utterance, if it is rejected the learner might be asked to try again. Experiments conducted so far indicated error rates of about 10%.

3.2.Error detection

In the DISCO system errors have to be detected in pronunciation, morphology, and syntax.

3.2.1.Pronunciation

In previous studies we investigated which pronunciation errors are made by learners of Dutch [8], and how these errors can be detected automatically [7]. For error detection, it has to be tested whether segments are present or not and whether they are realized correctly. This can be done by using confidence measures or similar classifiers at the segmental level. In our own studies we achieved accuracy scores between 82% and 94% [7] [9].

3.2.2.Syntax

While pronunciation error detection concerns detecting whether segments are realized or not, syntactic error detection

generally concerns detecting whether words are realized or not, and whether they are in the right order.

In phase 1, syntactically incorrect responses will be included in the list of predicted (correct and incorrect) responses (see 3.1). The output of phase 1 can thus be an incorrect utterance present in the predicted list. Additionally, in phase 2a, detailed analysis at word level might be carried out, e.g. confidence levels at word level to determine whether the correct words are present in the correct order.

3.2.3.Morphology

There are different types of morphological errors. Consider the following examples:
(c1) "gisteren maakte hij" (yesterday made he), i.e. the correct form is maakte, and incorrect are maak, maakt, maakten, which are all existing inflections of the root form "maken" (to make), but are not correct in the current context;
(c2) "gisteren ging hij" (yesterday went he), i.e. the correct form is "ging", and not other inflections of the root form "gaan" (to go): ga, gaat, gaan, *gaatte, etc.

Many morphological errors are similar to example c1, i.e. they concern segments that are inserted or deleted. Thus error detection boils down to detecting whether these segments are realized or not (especially /t/, /@/, and /n/). Other morphological errors are more similar to example c2. Therefore, the algorithms for detecting morphological errors will be a combination of the algorithms used for detecting pronunciation errors and those used for syntactic errors.

3.3.How to deal with technical limitations

Since ASR performance is not 100%, the feedback is likely to contain errors: false accepts (FA) and false rejects (FR). For phase 1 (speech recognition), a false accept means that what is recognized is not what was actually spoken: the learner gets feedback on something that was not said. A false reject means that an utterance is not recognized even though it is present in the predicted list of responses: the user will be asked to try again. For phase 2, a FA means that 'a form' is accepted although it is incorrect, and a FR means that it is rejected although correct. The modules will be first evaluated and optimized in isolation (see section 3) and later also in combination. After all, the modules are not independent. For instance, if in phase 1 a FA occurs, the detected errors can still be correct, e.g. if they concern errors in the correctly recognized part of the utterance. By varying thresholds, taking different points on the ROC curves, the influence (weights) of FA and FR in the different phases can be changed. In general, FRs are probably more confusing. But this may also differ from person to person, e.g. depend on the number and type of errors made. Possibility: adaptive weights. In any case, giving incorrect feedback should be avoided. However, if the thresholds are set too high, too conservative, in phase 1 the feedback often will be sth. like "try again", and in phase 2 there often will be no feedback on errors. It is clear that a careful balance should be found.

In order to limit the amount of confusion due to incorrect feedback, there are some other options. One is to show on the screen what is recognized, and thus the learner can see where the error detection is based on. Another possibility would be to ask for confirmation for every recognized utterance.

4. Conclusions

The results of the preparatory studies conducted so far have indicated how we can take account of the limitations of non-native ASR and still develop an application that is in line with current views on L2 learning and can support it through "some means of Focus on Form that is socially provided during meaningful communication and that recruits the learner's explicit conscious processing" [10].

5. Acknowledgements

The DISCO project is carried out within the STEVIN programme funded by the Dutch and Flemish Governments (http://taalunieversum.org/taal/technologie/stevin/).

6. References

[1] Swain, M., and Lapkin, S., "Problems in output and the cognitive processes they generate: A step towards second language learning", *Applied Linguistics*, vol. 16, pp. 371-391, 1995.

[2] Strik, H., "DISCO project website". Available: http://lands.let.kun.nl/~strik/research/DISCO [Accessed: May 15, 2009].

[3] Hubbard, P., "Interactive Participatory Dramas for Language Learning", Simulation and Gaming, vol. 33, pp. 210-216, 2002.

[4] Krueger, R.A. & Casey, M.A. (2000) Focus groups: a practical guide for applied research, California: Thousand Oaks.

[5] Engwall, O., and Bälter, O., "Pronunciation feedback from real and virtual language teachers", *Computer Assisted Language Learning*, vol. 20, no. 3, pp. 235-262, 2007.

[6] Menzel, W., Herron, D., Morton, R., Pezzotta, D., Bonaventura, P., and Howarth, P., "Interactive pronunciation training", *Re-CALL*, vol. 13, no. 1, pp. 67-78, 2000.

[7] Cucchiarini, C., Neri, A., and Strik, H., "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback", *Speech Communication*, to appear.

[8] Neri, A., Cucchiarini, C. and Strik, H., "Selecting segmental errors in L2 Dutch for optimal pronunciation training", *IRAL - International Review of Applied Linguistics*, vol. 44, pp. 357–404, 2006.

[9] Strik, H., Truong, K., de Wet, F. and Cucchiarini C., "Comparing different approaches for automatic pronunciation error detection", *Speech Communication*, to appear.

[10] Ellis, N.C., Bogart, P.S.H., 2007, Speech and Language Technology in Education: the perspective from SLA research and practice, Proceedings ISCA ITRW SLaTE, Farmington PA.