Automatic Speech Recognition in CALL systems: The essential role of adaptation

Helmer Strik, Joost van Doremalen, Catia Cucchiarini

1 Introduction

Recent advances in speech technology research coupled to the increasing importance of speaking proficiency as an essential skill to acquire by migrant workers learning a second language (L2) have led to a growing interest in developing CALL systems that make use of Automatic Speech Recognition (ASR) for practicing oral proficiency. Such systems offer L2 learners the opportunity of practicing speaking outside the language classroom, at their own pace, without time limitations, and in a stress-free environment. However, direct experience with ASR-based CALL applications reveals that developing effective systems is fraught with difficulties. One of the main challenges is adapting the technology in such a way that it can cope with the idiosyncrasies of non-native learner speech while at the same time detecting the errors with sufficient accuracy.

2 ASR-based CALL: the DISCO system

In this paper we address the problem of ASR adaptation within the framework of DISCO [1,2], a project aimed at developing a prototype of an ASR-based CALL application for learning Dutch as a second language (DL2). This 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. Such an application requires dedicated speech technology modules for recognition of non-native speech and error detection.

2.1 System adaptation

An ASR-based CALL system that has to provide corrective feedback on speech production will first of all have to determine what the learner is trying to say (speech recognition) before proceeding to an analysis of the form of the utterance (error detection). The first step of speech recognition may be very difficult in the case of non-native speakers, in particular those that are still in the process of learning the language. From research we known that non-native speech may differ from native speech with respect to pronunciation, morphology, syntax, and the lexicon. In general, the degree of deviation from native speech will be in inverse relation to the degree of proficiency in the target language. All these deviations make it very difficult to

H. Strik
Centre for Language and Speech Technology (CLST), Radboud University, Nijmegen
H.Strik@let.ru.nl

J. van Doremalen
Centre for Language and Speech Technology (CLST), Radboud University, Nijmegen
J.vanDoremalen@let.ru.nl

C. Cucchiarini
Centre for Language and Speech Technology (CLST), Radboud University, Nijmegen
C.Cucchiarini@let.ru.nl

recognize what a person is trying to say. To circumvent these problems measures have to be taken to adapt the recognizer to the idiosyncrasies of this sort of speech. These measures concern adaptation of the acoustic models, the language model and the lexicon, the main components of a speech recognizer. All these forms of adaptation will make the recognizer more tolerant with respect to the incoming speech so that even deviant realisations of the intended utterance will be recognized as valid attempts to produce a response. However, once an incoming utterance has been recognized as being an acceptable attempt at producing the response required, the speech recognizer has to go through the same utterance and carry out a different kind of analysis, a much stricter one, to determine whether the form of the utterance is correct.

These are exactly the challenges we face in the DISCO project. In DISCO we use a two-step procedure in which (1) first the content of the utterance is determined (what was said, speech recognition), and (2) subsequently the form of the utterance is analysed (how it was said, error detection). A common approach to limit the difficulties in speech recognition consists in applying techniques that restrict the search space and make the task easier. In line with this approach, in DISCO we combine strategies that are essentially aimed at constraining the output of the learner so that the speech becomes more predictable with techniques that are aimed at improving the decoding of non-native speech. This is achieved by generating a predefined list of possible (correct and incorrect) responses for each exercise.

2.2 Utterance selection and verification

Since learners thus have some freedom in formulating their responses, it first has to be determined which utterance (of the predefined list) was spoken, which is done by means of utterance selection. There is always the possibility that the learner's response is not present in the predefined list or that utterance selection does not select the correct utterance from the list. To check this, utterance verification is carried out. In the first step of the two-step procedure, two phases can thus be distinguished, (1a) utterance selection, and (1b) utterance verification (UV). Experiments conducted so far indicated that reasonable levels of accuracy could be obtained at the stage of (1a) utterance selection (about 8-10%) and (1b) utterance verification (10%) [3].

2.3 Error detection

In the phase of error detection different approaches are adopted for pronunciation, morphology and syntax. Syntactic errors can be addressed at stage 1 by including incorrect responses in the list of predicted (correct and incorrect) responses. The output of stage 1 can thus be an incorrect utterance present in the predicted list. Additionally, 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.

For pronunciation 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. For the well-known goodness of pronunciation (GOP) algorithm we obtained accuracy scores of 82-89% [4,5]. We also evaluated classifiers in which acoustic phonetic features were used. With these classifiers even better results were obtained, i.e. accuracy scores of about 90-94% [5].

Many morphological errors have to do with whether segments are present or not and for these cases error detection is very similar to pronunciation error detection. Other morphological errors are not limited to the presence or absence of a segment but concern multiple aspects of a word. In other words, the algorithms used for detecting morphological errors will be a combination of those used for detecting pronunciation errors and those used for syntactic errors.

2.4 Remedial exercises

Achieving sufficient accuracy in the stages of speech recognition and error detection is essential to be able to provide useful corrective feedback. In addition, in the DISCO project the results of error detection will be used as guidelines to assign remedial exercises to the learner. This is of course an additional form of adaptation that is envisaged in the DISCO system and that is highly dependent on the performance achieved by the speech technology modules.
In the presentation all these issues will be discussed in detail.

References

1. H. Strik , F. Cornillie, J. Colpaert, J. van Doremalen, C. Cucchiarini, 'Developing a CALL System for Practicing Oral Proficiency: How to Design for Speech Technology, Pedagogy and Learners', Proceedings of the SLaTE-2009 workshop, Warwickshire, England (2009)
2. http://lands.let.ru.nl/~strik/research/DISCO/
3. J. van Doremalen, H. Strik, C. Cucchiarini, 'Utterance Verifiction in Language Learning Applications', Proceedings of the SLaTE-2009 workshop, Warwickshire, England (2009)
4. S. Kanters, C. Cucchiarini, H. Strik, 'The Goodness of Pronunciation Algorithm: a Detailed Performance Study', Proceedings of the SLaTE-2009 workshop, Warwickshire, England (2009)
5. H. Strik, K. Truong, F. de Wet, C. Cucchiarini, 'Comparing different approaches for automatic pronunciation error detection', Speech Communication, Volume 51, Issue 10, October 2009, Pages 845-852 (2009)