Speech technology for language tutoring

Helmer Strik [1] *, Ambra Neri* [1] *, and Catia Cucchiarini* [1]

[1] Department of Language and Speech, Radboud University Nijmegen, The Netherlands

h.strik@let.ru.nl, a.neri@let.ru.nl, c.cucchiarini@let.ru.nl

Abstract

Language learners are known to perform best in one-on-one interactive situations in which they receive optimal corrective feedback. However, one-on-one tutoring by trained language instructors is costly and therefore not feasible for the majority of language learners. This particularly applies to oral proficiency, which requires intensive tutoring. Computer Assisted Language Learning (CALL) systems that make use of Automatic Speech Recognition (ASR) seem to offer new perspectives for language tutoring. In this paper we explain how.

Index Terms: Computer Assisted Language Learning (CALL), Automatic Speech Recognition (ASR), language tutoring.

1. Introduction

Language learners are known to perform best in one-on-one interactive situations in which they receive optimal corrective feedback. The two sigma benefit demonstrated by Bloom [2] has provided further support for the advantages of one-on-one tutoring relative to classroom instruction. However, one-on-one tutoring by trained language instructors is costly and therefore not feasible for the majority of language learners. In the classroom, providing individual corrective feedback is not always possible, mainly due to lack of time. This particularly applies to oral proficiency, where corrective feedback has to be provided immediately after the utterance has been spoken, thus making it even more difficult to provide sufficient practice in the classroom.

The emergence of CALL systems that make use of Automatic Speech Recognition (ASR) seems to offer new perspectives for language tutoring. These systems can offer extra learning time and material, specific feedback on individual errors and the possibility to simulate realistic interaction in a private and stress-free environment.

At the same time the increasing mobility in Europe and in the world at large together with the recent emphasis on promoting plurilingualism and linguistic diversity in Europe has led to a situation in many countries in which the demand for language lessons outstrips supply. As a consequence, new methods and technologies that make language learning more efficient and effective are called for.

ASR-based CALL could be employed to develop new methods for teaching literacy, reading, oral proficiency, speaking fluency, and vocabulary. In this paper we first review some studies that have employed ASR for language learning with mixed results. We then go on to consider important aspects of software design and a number of technological challenges. Finally, we draw some conclusions and consider challenges and opportunities for the future.

2. CALL applications

Speech technology is already used in several CALL applications. However, some researchers are skeptical about the usefulness and effectiveness of ASR-based CALL programs: evidence gathered in different lines of research seems to confirm that either speech technology is not mature enough, or ASR-based CALL programs are not effective in improving second language (L2) skills [e.g. 3, 5]. For the sake of our own research, we have studied this literature thoroughly and have gradually acquired the impression that, while it is undeniable that speech technology still presents a number of limitations, especially when applied to non-native speech, part of this pessimism is in fact due to misconceptions about this technology and CALL in general.

2.1. CALL & ASR

ASR dictation packages are being used by a growing number of people working in different branches. These packages offer good performance at a reasonable price and are readily available. It is probably for these reasons that some teachers and CALL practitioners have become interested in these programs as a possible tool to teach L2 skills [3, 5].

Derwing, Munro, and Carbonaro [5] investigated the usefulness of ASR for CALL by evaluating the performance of a standard dictation package, Dragon NaturallySpeaking Preferred, in identifying pronunciation errors in the L2 speech of Cantonese and Spanish learners of English. The authors propose two criteria for establishing the effectiveness of ASR in providing corrective feedback on L2 speech errors. First, the software should be able to recognize the oral language of 'English as a Second Language' (ESL) speakers at an acceptable level. Second, the software's identification of L2 speech errors must resemble that of native, human listeners.

On the basis of their study, Derwing et al. [5] conclude that ASR "cannot be considered to be of benefit to ESL speakers" [5: p. 602], that "the computer's output might be confusing to ESL students" and that "the observed levels would frustrate a user hoping for reliable feedback on intelligibility" [5: p. 600]. However, it is important to stress that the first conclusion does not apply to ASR in general, but to the specific ASR dictation package tested in this study, which was never intended for L2 learning. Analogously, the second and third conclusions are based on the incorrect assumption that the output of a dictation system can be used as a basis for providing feedback to L2 learners. Although the authors clearly state what the domain of their evaluation is in the introduction, they fail to relate their negative results to the characteristics of the specific technology they used, which may lead many to generalize those conclusions to the use of speech technology as a whole for L2 training.

Coniam [3] conducted a study aimed at exploring "the potential of the use of voice recognition technology with second language speakers of English" [3: p. 49] by testing the dictation package Dragon NaturallySpeaking on ten native speakers and ten Cantonese speakers of English. The recognition accuracy of the system was examined for both speakers groups for an excerpt read from a book. Besides, the author compared the output of the recognizer for native and non-native speakers with the original text in an attempt to

identify phonological patterns, e.g. sound substitutions in the speech of the Cantonese speakers, on the basis of the CSR output. The results show that the system's accuracy is higher for native speech than for non-native speech and Coniam [3] concludes that "voice recognition technology is still at an early stage of development in terms of accuracy and single-speaker dependency" [3: p. 49] although it might have potential in the future.

In these studies unsatisfactory results are obtained when standard dictation systems are used for CALL. But dictation systems are not suitable for L2 training, CALL requires dedicated speech technology. Apart from the fact that current dictation packages are usually developed for native speakers, the major problem in using this technology for CALL has to do with the fact that dictation and CALL have different goals which require different approaches in ASR. The aim of a dictation package is to convert an acoustic signal into a string of words and not to identify L2 errors, which requires a more complex procedure. Consequently, the negative conclusions should be related to this specific case and not to speech technology in general.

2.2. Software design

ASR-based CALL systems can recognize what a student actually uttered, to detect errors, and to provide immediate feedback on them. However, the technology needed for such systems is highly complex and still has a number of important limitations that should seriously be reckoned with when designing applications for L2 pronunciation error detection [see e.g. 6, 7].

Another important aspect of system design is pedagogical guidelines. Many commercial CALL systems present fancy looking features that are likely to impress the buyers, but that in fact do not serve any real pedagogical purpose, thus they do not meet the real needs of L2 learners The design of these systems seems to be driven more by a technology push, rather than being based on a comprehensive analysis of the requirements that the system must meet to be effective and efficient. This may in part be due to a difficulty involving different experts in the design phase of a CALL system, or more fundamentally, to the absence of clear pedagogical guidelines that suit CALL. Here we present examples of how inadequacies in system design leading to disappointing performance often end up being unjustly attributed to speech technology.

A fine example of how speech technology can be employed to diagnose segmental errors and provide feedback on them is provided by the ISLE (Interactive Spoken Language Education) project [9, 11]. Within the framework of this project, a system for German and Italian learners of English was developed which provided feedback on pronunciation errors at phoneme and word-stress level. The feedback on phonemes consists in highlighting the grapheme corresponding to the erroneous phoneme in the utterance and by showing on the screen both a frequent word containing the correct target phoneme and another frequent word containing the student's incorrect realization of it. The student can listen to both sounds in a focused way and try to notice the differences. A list of words containing the problematic target phoneme is optionally provided for training.

This system is therefore not only able to determine where a segmental error occurred in an utterance, but also what sound was realized. This approach is based on the belief that a CALL system should not only indicate that there is an error, but also specify where the error is located and how it should be corrected [11]. While this design seems almost ideal, the performance of the system is poor. The authors report that only 25% of the errors are detected by the system and that over 5% of correct phones are incorrectly classified as errors, whereby "students will more frequently be given erroneous discouraging feedback than they will be given helpful diagnoses" [11, p. 54].

However, performance could probably be improved by adopting a slightly different design that takes more account of the limitations of speech technology. Given these limitations, the ISLE system is likely to make mistakes at various stages: in recognizing the utterance, in locating the error, in diagnosing the problem, and thus also in presenting the example words. A slightly less ambitious system that has to make fewer decisions is also likely to make fewer errors. For example, a system that only indicates the part of a word or utterance that was mispronounced, without indicating exactly which erroneous sounds it recognized would be less sophisticated and, probably, less error-prone. Although it is more desirable to provide diagnostic information in Computer Assisted Pronunciation Training (CAPT), we have to conclude that such a system cannot guarantee a satisfactory performance yet with current technology. But, as it turns out, only showing mispronunciations might be sufficient feedback, and a CAPT system that does this may just as well be useful for improving pronunciation as we have shown in our research [12, 13].

Another example concerns some commercial systems which seem driven by technological innovations rather than by pedagogical guidelines. In some of these systems, that advertise themselves by mentioning that ASR is used, the feedback consists of an overall score and a graphical display with waveforms or spectrogram's, usually one window displaying the utterance spoken by the language learner, and another window with a reference utterance. These graphical displays look like an invitation for the student to try to understand how the two are related, but practice with the programs generally does not shed light on this aspect [1, 12]. The student may eventually end up realizing that there is no relation between the two [1, 12], which impoverishes the pedagogical value of this kind of feedback. In addition, the fact that the system shows two comparable displays, one representing the student's utterance and one representing the model utterance, wrongly suggests that the student should produce an utterance that closely corresponds to that of the model. In fact, this is not necessary at all: two utterances with the same content may both be very well pronounced and still have waveforms or spectrograms that are very different from each other. Moreover, waveforms and spectrograms are not easily interpretable for students [1, 12]. Even students with knowledge of acoustic phonetics are likely to find it hard to extract the information needed to improve pronunciation from these displays, since there is no simple correspondence between the articulatory gesture and the acoustic structure in the properties displayed. As many authors have observed, this type of feedback is not easy to comprehend and thus of limited pedagogical value [6, 7, 11, 12]. Despite their little pedagogical value, it might be that these programs are flashy and impressive [1, 12], a factor whose importance should not be underestimated in commercial products.

In these cases, shortcomings in the design of the ASR-based CALL programs contribute to creating the impression that speech technology is to blame. One of the reasons why these systems perform poorly is that they were designed without taking due account of the limitations of speech technology: as a result, programs that are too ambitious given the state of the art of this technology are developed. When

performance turns out to be disappointing, one then gets the impression that ASR as a whole is inadequate for the purpose of automatic training of L2 learners, while a better design with that very same technology might have produced a more effective product. This might in part explain why these systems eventually turn out to be ineffective in improving L2 pronunciation, which then helps create the impression that the disappointing learning results are due to the inadequacy of speech technology for this specific purpose.

3. Technological challenges

In the previous section we have provided some examples of using speech technology in CALL applications for which the results were disappointing. We explained that these disappointing results are to a large extent the result of a combination of factors, not all of which are related to the inadequacy of speech technology. An important factor is speech recognition, but this is only one part of a complex system, for which complex decisions have to be taken. These decisions might not always lead to the desired learning outcomes, as we have seen.

In this section, we intend to show how speech technology, despite its undoubted limitations, can still be employed in a meaningful and useful way in CALL applications. The challenges are the following. We should obviously try to improve the technology. However, since these improvements are likely to be gradual, as they have been during the last decades, we should also try to make optimal use of current technology, taking into account what is possible and what isn't possible with current technology.

3.1. Speech recognition

In the ASR community, it has long been known that the differences between native and non-native speech are so extensive as to degrade ASR performance considerably [9, 15, 16]. ASR-technology is sensitive to the degree of mismatch between acoustic models and incoming speech, to vocabulary size, and to task complexity. Still, there are some ways in which ASR performance can be improved.

Instead of using an ASR that is trained only on native (L1) speech, an ASR should be used which is trained on non-native (L2) speech (possibly in combination with L1 speech): lexica with non-native pronunciation networks, language models based on words and word orders as spoken by non-natives, and acoustic models that represent the way non-natives pronounce sounds. For the acoustic models there are several possibilities: simply train them on L2 speech [8], use acoustic models of L1 and L2 in parallel [10], or a combination of L1 and L2 models [8, 16], and, optionally, also include intermediate phones.

Besides improving the ASR, one could also make the task of the ASR less difficult. For a constrained task, it is possible to use a constrained lexicon and language model, which will increase the performance. If the number of possible answers is limited, one could even use utterance verification techniques (e.g. confidence measures) to select the utterance that was spoken from the list. In that way, performance could even be improved more. The challenge then is to develop engaging items, for which the possible answers can be predicted. Eliciting speech from the speakers in such a way that the lexicon is known in advance [6, 9] can be achieved in various ways, for instance by asking closed-response questions instead of open-response questions, by having speakers read aloud sentences, or repeat auditorily primed sentences. The choice of the strategy will depend on the type of application.

Such a strategy has been employed in a CAPT system, called Dutch-CAPT, we developed for learning Dutch [12, 13]. This system was based on an already existing multimedia learning system, called 'Nieuwe Buren' ('new neighbors'), which did not make use of speech technology. We selected some of the lessons, added speech technology to it which made it possible to give feedback on the spoken utterances, and developed a new user interface (Fig. 1). Each lesson started with watching a video, followed by some exercises: role-playing (Fig. 1), question answering, and reading. For all items, utterance verification was used for recognition.

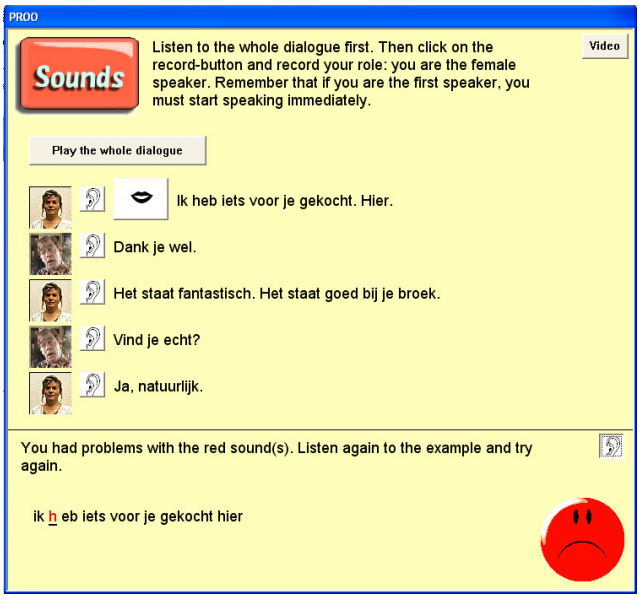

Figure 1. A screen shot of the Dutch-CAPT system.

Our design was based on a thorough study of existing call systems. For instance, we decided to use feedback that is not as detailed as the feedback given in the ISLE system (see section 2.2). In this way we managed to reduce the number of errors in the feedback. An example of the interface of our system is shown in Figure 1: an indication is given of which sounds are not pronounced correctly (red and underlined), and the language learner has the opportunity of listening to his utterance, an example utterance, and try again. When the system was tested, it turned out that language learners who used this system only four times for about 30 to 60 minutes improved more than a control group that did not use the system [12, 13].

3.2. Assessment

Once the speech signal has been recognized, it has to be evaluated. This phase rests on the opposite assumption as that underlying the recognition phase: while good recognition of the students' speech implies that the system be tolerant of discrepancies between the incoming speech and the native speech models, good scoring requires the system to look exactly for those discrepancies. Different terms have been used by the various authors for this stage: pronunciation scoring, pronunciation grading, error detection, error localization, error identification etc. Although these terms are often used interchangeably, they can be used to refer to two different activities, as will be explained below.

In general, error detection indicates the procedure by which a score at a local (phoneme) level is calculated, while pronunciation grading stands for the procedure that is followed

to calculate a global score at the utterance level, which could also be a weighted average of the local scores. Seen in this light, error detection can be considered a specific case of pronunciation grading, but there is more to it than meets the eye. In fact, error detection and pronunciation grading can be viewed as two different tasks, with a different goal and a different output. For grading, more global measures can be used, such as temporal measures [4]. The relation between human and automatic grading becomes better if longer stretches of speech are used, i.e. complete utterances or a couple of utterances.

Error detection can also be carried out for number of utterances in combination; however, for language learning one generally prefers immediate feedback (and not feedback after a number of utterances have been spoken). For pronunciation error detection, some approaches can be used:

1. focus on frequent errors
2. ASR-based metrics
3. acoustic phonetic classifiers

In the first approach, errors frequently made by language learners are explicitly taken into account [10]. For instance, if the sound /r/ is often pronounced as /l/ (e.g. 'angly' instead of 'angry'), then this frequent substitution can be hard-wired in the pronunciation models. If the speech recognizer decides that the best path 'goes through' the /l/ sound, then the system knows that probably this pronunciation error is made.

In the second approach, ASR-based metrics are used, such as posterior probabilities and (log) likelihoods ratios [8 9]. Previous research has shown that these confidence measures can be used for detecting pronunciation errors [8, 9, 16]. A special case concerns the so-called goodness of pronunciation algorithm (GOP) [16], which has been used quite often. We also studied the GOP algorithm in our research, and used it in our Dutch-CAPT system. If trained well, the GOP algorithm works quite well: on average about 80% of the sounds are classified correctly. However, there are large variations between persons and sounds. If specific settings could be used for each person sound combination, better results could be achieved; but currently this is not possible in practice. The challenge here is to find groups for which similar settings perform well.

Acoustic phonetic classifiers are not often used in call applications; still they can be useful [14]. We compared the results of acoustic phonetic classifiers to those obtained with the GOP algorithm, and it turned out that results for acoustic phonetic classifiers were better [14].

As often, a combination of approaches probably will yield the best results. Therefore, the challenge here is to find the proper combination of approaches and settings for which the results are best.

4. Future: challenges and opportunities

In this paper we have shown that ASR, in spite of its limitations, already holds great potential for language tutoring and could be employed for various language learning goals such as literacy development, reading, oral proficiency, and vocabulary learning. It is clear that developing good applications requires mixed expertise: knowledge of speech technology, education / pedagogy, language acquisition, software design and development. Developing good products therefore requires that the right people work together: speech technologists, teaching professionals, software designers and industrial partners (e.g. publishers).

At the same time, a globalized world characterized by increasing internationalization and mobility will continue to require products and material that make it possible to learn new languages efficiently and effectively to guarantee the integration of migrant workers into their new surroundings.

5. References

The references below are listed in alphabetical order.

[1] ALR (1998) Putting Pronunciation Programs Through Their Paces, American Language Review, 2. http://www.languagemagazine.com/internetedition/mj98/eepp56.html (retrieved November 29, 2007)

[2] Bloom, B. S. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 1984, 4-16.

[3] Coniam, D. (1999) Voice recognition software accuracy with second language speakers of English. System, 27, 49-64.

[4] Cucchiarini, C., Strik, H., & Boves, L. (2002) Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. Journal of the Acoustical Society of America, 111(6), 2862-2873.

[5] Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? TESOL Quarterly, 34, 592-603.

[6] Ehsani, F., & Knodt, E. (1998) Speech technology in computer-aided learning: Strengths and limitations of a new CALL paradigm. Language Learning & Technology, 2, 45-60.

[7] Eskenazi, M. (1999) Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype, Language Learning & Technology, 2, 62-76.

[8] Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000b) Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30, 121-130.

[9] ISLE 1.4 (1999) Pronunciation training: Requirements and solutions, ISLE Deliverable 1.4. Retrieved February 27, 2002, from http://nats-www.informatik.uni-hamburg.de/~isle/public/D14/D14.html.

[10] Kawai, G., & Hirose, K. (1998) A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training, Proceedings of ICSLP, Sydney, Australia, 1823-1826.

[11] Menzel, W., Herron, D., Bonaventura, P., & Morton, R. (2000) Automatic detection and correction of non-native English pronunciations, Proceedings of InSTiL, Dundee, Scotland, 49-56.

[12] Neri, A. (2007) The pedagogical effectiveness of ASR-based computer assisted pronunciation training. PhD thesis, University Nijmegen.

[13] Neri, A. Cucchiarini, C. and Strik, H. (2007) Pronunciation training in Dutch as a second language on the basis of automatic speech recognition, Stem, Spraak- en Taalpathologie, 159-169.

[14] Strik, H., Truong, K., De Wet, F., and Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. Proceedings of Interspeech 2007, Antwerp.

[15] Van Compernolle, D. (2001) Recognizing speech of goats, wolves, sheep and ... non-natives. Speech Communication, 35, 71-79.

[16] Witt, S. (1999) Use of Speech Recognition in Computer Assisted Language Learning. PhD thesis, University of Cambridge.