Automatic Speech Recognition for second language learning: How and why it actually works

Ambra Neri, Catia Cucchiarini, and Wilhelmus Strik
A²RT, Department of Languge and Speech, University of Nijmegen, The Netherlands
{A.Neri, C.Cucchiarini, W.Strik}@let.kun.nl

ABSTRACT

In this paper, we examine various studies and reviews on the usability of Automatic Speech Recognition (ASR) technology as a tool to train pronunciation in the second language (L2). We show that part of the criticism that has been addressed to this technology is not warranted, being rather the result of limited familiarity with ASR technology and with broader Computer Assisted Language Learning (CALL) courseware design matters. In our analysis we also consider actual problems of state-of-the-art ASR technology, with a view to indicating how ASR can be employed to develop courseware that is both pedagogically sound and reliable.

1. INTRODUCTION

CALL systems have become popular tools to train pronunciation in the L2 because they offer extra learning time and material as well as the possibility to practise in a stress-free environment. With the integration of ASR technology, these systems, which we will refer to as CAPT (Computer Assisted Pronunciation Training) systems, can even provide limited interaction: the computer understands the student's speech and reacts accordingly, thus making the learning process more realistic and engaging, and can provide feedback on the quality of the student's speech. While students generally enjoy learning with speech-enabled systems, a number of researchers and educators are sceptical about the usability of ASR for pronunciation training in the L2, because this technology still suffers from a number of limitations.

For this reason, several attempts have been made to establish the effectiveness of ASR technology for CAPT. In many publications that have appeared in the language teaching community, criticism has been expressed with regard to the two main features of ASR that is to be used by language learners: the ability to recognize accented or mispronounced speech, and the ability to provide meaningful evaluation of pronunciation quality. In particular, it seems that criteria such as those proposed in [1] that a) recognition performance must be at an acceptable level and that b) the identification of L2 speech errors must resemble that of native listeners in many cases are not met.

A thorough study of this literature and of the specialized literature on speech technology has nevertheless convinced us that this pessimism is not entirely justified. While it is undeniable that state-of-the-art ASR still presents a number of problems, we believe that some of the problems reported in the publications we examined are actually due to factors that are not directly related to ASR technology, but were attributed to it because of little familiarity with ASR technology and with design matters within ASR-based CAPT. In the following sections we consider the problems described with a view to establishing which ones are really due to limitations in the technology and to explaining how, in spite of these limitations, ASR can be used to develop systems that are robust enough to handle non-native speech, and that, at the same time, are able to meet sound pedagogical criteria.

2. AUTOMATIC SPEECH RECOGNITION FOR CAPT

The ideal ASR-based CAPT system can be described as a sequence of five phases, the first four of which strictly concern ASR components that are not visible to the user, while the fifth has to do with broader design and graphical user interface issues.

1) Speech recognition: the ASR engine translates the incoming speech signal into a sequence of words on the basis of internal phonetic and syntactic models. This is the first and most important phase, as the subsequent phases depend on the accuracy of this one. Besides, this phase alone already allows devising a range of computer-based activities to train communicative skills in the L2, such as interactive dialogues with the computer and speech-enabled multiple-choice exercises. However, the main pedagogical advantage that ASR-based CAPT can offer for training oral skills in the L2 is the provision of an evaluation of pronunciation quality. The following phases show how this evaluation is possible.
2) Scoring: this phase makes it possible to provide a first, global evaluation of pronunciation quality in the form of a score. The ASR system analyses the spoken utterance that has been previously recognized. The analysis can be done on the basis of a comparison between temporal properties (e.g. rate of speech) and/or acoustic properties of the student's utterance on one side, and natives' reference properties on the other side: the closer the student's utterance comes to the native models used as reference, the higher the score will be. The usefulness of automatic

ISBN 1-876346-48-5

scoring for pronunciation training is evident, as it gives the learner immediate information on overall output quality and on how this can improve over successive attempts.
3) Error detection: the system can locate the errors in the utterance and indicate to the learner where s/he made mistakes. This is generally done on the basis of so-called confidence scores that represent the degree of certainty of the ASR system that the recognized individual phones within an utterance actually match the stored native models used as a reference. Signalling that a certain sound within a word is problematic can be particularly useful to raise awareness in the learner of that problem and thus help her/him to focus and practise more on that area.
4) Error diagnosis: the ASR system identifies the specific *type* of error that was made by the student and suggests how to improve it, because a learner may not be able to identify the exact nature of his pronunciation problem alone. This can be done by resorting to previously stored models of typical errors that are made by non-native speakers.
5) Feedback presentation: this phase consists in presenting the information obtained during phases 2,3, and 4 to the student. It should be clear that while this phase implies manipulating the various calculations made by the ASR system, the decisions that have to be taken here – e.g. presenting the overall score as a graded bar, or as a number on a given scale – have to do with design, rather than with the technological implementation of the ASR system. This phase is fundamental because the learner will only be able to benefit from all the information obtained by means of ASR if this is presented in a meaningful way.

3. PROBLEMS REPORTED

Developing ASR software for CAPT systems is a complex job that would ideally require software developers, speech technologists, and educators to work together. For obvious reasons, evaluations of these systems that are based on this whole range of expertise, that are objective and accessible to teachers -probably those who are most interested in the effectiveness of these applications - are extremely rare. We found that several evaluations addressed to teachers or generic CALL practitioners tended to be flawed by little familiarity with ASR technology and design issues. In the following sections we deal with the problems reported, by locating them at the various phases listed above, with a view to verifying their actual nature and cause.

3.1 Speech recognition phase

The first and most important task of a speech recogniser is to recognize speech correctly. No teacher will want his students to work with a system that does not guarantee acceptable levels of recognition accuracy [1]. For this reason, [1] and [2] evaluated the recognition performance of ASR software in two standard dictation packages for English. Both studies found that, while it offers good results for native (English) speakers (90% accuracy), this software performs less well for non-native (Cantonese and Spanish) speakers and is therefore not yet mature for being used in the L2 classroom. The problem with these studies is that they indirectly suggest that ASR technology as a whole is not reliable enough to be used for CAPT, whereas a dictation package, like the ones used in these studies, is very different from a CAPT system, among other things because it targets a different group of users, i.e. native speakers. This implies that the ASR technology that is employed within these two types of systems is different.

ASR that is developed to recognize native speech, as in dictation packages, is known to perform poorly on non-native speech because of different acoustic properties in the speech of non-native speakers. Since the first studies appeared [3,4], speech technologists have developed a number of measures to ensure that ASR of non-native speech achieves acceptable levels of performance and have applied them successfully to tune specific ASR engines for the task of recognizing non-native speech in CALL and CAPT courseware [5,6,7]. Commonly-used methods consist in training the ASR engine with speech from both native and non-native speakers, adapting the native models that are to be used as a reference to non-native acoustic properties, and possibly adding to the ASR phone inventory models of phones that are not present in the target language, but that are likely to be produced by non-native speakers.

Besides, a rule of thumb to ensure good recognition performance is to keep the recognition task as simple and as limited as possible, by carefully designing the learning activities. Recognizing one sentence out of a list of three phonetically different responses is much easier than recognizing one sentence spoken into a standard dictation package that refers to a dictionary of thousands of words and word combinations [5,6]. Most ASR-based activities in current CAPT systems are good examples of this strategy (e.g. [8]), and indeed, most teachers seem satisfied with the recognition performance of non-native speech within these systems.

3.2 Evaluation phase: scoring, error detection, error diagnosis and feedback presentation

As already mentioned, the other important task for ASR within CAPT is to evaluate the student's speech correctly, possibly in a way that is comparable to human judgments [1]. The first necessary condition to provide valid information on the student's speech is accurate recognition. As we just saw, if the right type of ASR technology is used and the recognition task is designed carefully, we can ensure acceptable levels of recognition accuracy. Another condition that needs to be met concerns the scoring phase: automatic scores on pronunciation quality must resemble human judgments of the same speech sample. A number of teachers and researchers who evaluated ASR-based comprehensive CALL systems and CAPT systems have reported problems with this phase. Reesner [9] evaluated *Tell Me More French*, a comprehensive CALL system that provides an overall pronunciation score and instantaneously displays the waveform of the student's speech together with the waveform of the model utterance pronounced by a native speaker. Reesner observed that

ISBN 1-876346-48-5

while the combined presentation of overall score and graphical displays seems to imply that the two are somehow related to each other, neither he nor his students were able to find any clear relationship. Zheng [10], in his turn, comments that the feedback in the Chinese version of the same program is confusing, and he seems to attribute this to a problem with the 'speech recognition algorithm' (p.4). More specifically, Zheng claims that it is very difficult for the students to modify their pronunciation so as to match the model waveform. This is not surprising: while the simultaneous display of the two waveforms in this system may very well be taken as an invitation to produce utterances whose waveform closely corresponds to that of the models, this is not the real purpose of pronunciation training. Two utterances with the same content may both be very well pronounced and still have waveforms that are very different from each other. Many researchers have expressed doubts on the pedagogical value of these types of displays for this reason [11,12]. Besides, even a trained phonetician would find it difficult to extract information to correct one's pronunciation from these displays. Zheng's point deserves serious consideration because it might explain why certain CAPT systems that use this kind of feedback do not always turn out to be effective in improving L2 pronunciation (see [13]). However, it should be noted that this problem has nothing to do with the speech recognition algorithm: in fact, it is not even necessary to resort to ASR technology in order to produce this type of displays. If those displays are available in a program, it is simply because of a choice made by the developers (possibly guided by marketing experts who consider technological innovations paramount to pedagogical requirements [11]). Rather than to the scoring phase, this problem belongs to the 'feedback presentation' phase.

The automatic score provided by means of ASR has been object of further criticism. In [14], Wildner reported that native speakers sometimes received lower ASR-based scores than non-native speakers. Similarly, [15] found the scoring algorithm in *TriplePlayPlus!* inaccurate for the students at the more advanced level: utterances where the final syllables had been left out were deemed acceptable by the ASR system. [12] observed that the automatic ratings provided by means of ASR in *Tell Me More Japanese* sometimes differed from the teacher's ratings of the same utterances. Reesner [9] found the same type of score in the French version to be little informative and had troubles identifying its basis. These problems may be due to different specific causes, as the various ASR systems used may have been developed with different techniques, but all symptoms point to a serious problem with scoring. As a matter of fact, this is by no means a solved issue. Speech technologists are still trying to find the best measures with which to provide a meaningful score: these should be based on specific pronunciation aspects on which the student can work (but which are difficult to capture automatically) and, at the same time, they should result in a score that is similar to that provided by human listeners. Temporal measures, for instance, are strongly correlated with human ratings of pronunciation and fluency [16,17], which means that they are able to provide reliable scores for both native and non-native pronunciation assessment, but not necessarily for pronunciation training. A low score based on these measures would imply that the student should speak faster, or make fewer pauses, an indication that has little pedagogical value. It is thus necessary to integrate this kind of global evaluation by more detailed evaluation of specific problems.

Error detection represents a step further in the degree of detail of evaluation. Of course, in order for the evaluation to be meaningful, this phase should correctly locate possible pronunciation errors. Some of the systems that have been evaluated do not provide this type of information.[1] The more advanced systems that do include this phase, like [8], have generally received positive comments by the evaluators. This, together with studies from the speech technology field [17], seems to indicate that, by using the right combination of scores, segmental errors can be detected with reasonable accuracy.

Another condition that should be met in order to provide meaningful, human-like feedback concerns error diagnosis: ideally, a system should be able to provide a detailed diagnosis of a pronunciation problem and suggest the appropriate remedial steps, just like an ideal human tutor. With regard to this aspect, Hincks [13] complains about the inability of many commercial systems to diagnose specific problems and give corrective feedback rather than evaluative feedback. According to Hincks, the latter type of feedback would be more effective from a pedagogical point of view, especially for the more advanced learners. Recent research on ASR-based CAPT has nevertheless shown that this technology is not yet mature to provide reliable detailed diagnoses of pronunciation errors [18]. Moreover, Hinks's hypotheses are based on a study of *Tell Me More,* a program that makes use of waveforms to provide feedback, of which we have already discussed the dubious pedagogical value. Furthermore, research on corrective feedback does not corroborate Hincks' views. For instance, Lyster [19] found that recasts, i.e. the correct repetition of a mispronounced utterance without any further explanation, as in most teacher-to-student interactions, might be sufficient to correct deviant pronunciation in the short term. Similarly, [20] hypothesize that detailed feedback might not be necessary for proficient learners, who are already familiar with the linguistic inventory (e.g. correct sounds of the target language) and only need to be directed to the correct alternative when they make a mistake.

Finally, providing meaningful feedback means providing feedback that can be interpreted by the learner. This implies that all the information obtained in the first phases by the

[1] Contrary to what some researchers have attempted (see [1]), these systems should not be used for the purpose of evaluating automatic error detection, because the ASR software they contain was not conceived for providing feedback on pronunciation quality and is obviously bound to fail in this task.

 ISBN 1-876346-48-5

ASR system needs to be processed in the last phase, together with information obtained by other possible sources, and presented to the student in a clear, unambiguous way. How this is done is a matter of design and bears no relation to the pure technological aspects of the ASR engine. As we have seen, many researchers have expressed criticism on the use of certain feedback forms, such as waveforms, which has led some to unjustly blame ASR technology.

4. CONCLUSIONS

This analysis of various reviews on ASR technology for CAPT has revealed that part of the criticism addressed to this specific type of technology is not pertinent, but is rather due to a limited knowledge of ASR technology and of design issues. In spite of some limitations, we have seen that ASR technology can be employed to develop systems that can recognize and evaluate non-native speech in a way that resembles that of native listeners.

First, if the appropriate software is used and the speech-enabled learning activities are designed carefully, acceptable levels of recognition performance can be reached for non-native speech. Second, provided the right measures are employed, it is possible to provide human-like scores on overall pronunciation quality, but because the measures that are currently available cannot be used alone as a basis to provide information on specific pronunciation problems, the score they yield should be integrated with error detection, a task that appears to be performed satisfactorily. With regard to error diagnosis, given the limited reliability of ASR-based feedback at this phase and the scarcity of systematic studies on the effectiveness of various types of feedback, it seems safer to let the students practice with CAPT systems that are not too ambitious, but that can guarantee correct feedback in the majority of the cases. Finally, the success of these phases does not solely depend on the technical implementation of the ASR system, but relies to a large extent on the way the feedback is presented.

REFERENCES

[1] T.M. Derwing, M.J. Munro, and M. Carbonaro, "Does popular speech recognition software work with ESL speech?", *TESOL Quarterly* 34, 592-603, 2000.

[2] D. Coniam, "Voice recognition software accuracy with second language speakers of English", *System* 27, 49-64, 1999.

[3] L.M. Arslan, J.H.L. Hansen, "Language Accent Classification in American English", *Speech Communication* 18, 353-367, 1996.

[4] V. Digalakis, L. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods" *IEEE Transactions Speech and Audio Processing,* 294-300, 1996.

[5] ISLE D3.3 (1999) "Recognition of learner speech", ISLE Deliverable D3.3.

[6] H. Franco et al., "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning", *Proc. of InSTIL*, Scotland, 123-128, 2000.

[7] S.M. Witt, S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning", *Speech Communication* 30, 95-108, 2000.

[8] Auralog (2000) *Tell me More - User's Manual*, Montigny-le-Bretonneux, France.

[9] T. Reesner, "Tell Me More French", Software review, *CALICO Journal*, 19, 419-428, 2002.

[10] T. Zheng, "Tell Me More Chinese", http://www.calico.org/CALICO_Review/review/tmm-chinese00.htm

[11] A. Neri, C. Cucchiarini, H. Strik, and L. Boves "The pedagogy-technology interface in Computer Assisted Pronunciation Training", *Computer Assisted Language Learning* 15, 441-447, 2002.

[12] K. Miura, "Tell Me More Japanese", http://www.calico.org/CALICO_Review/review/tmm-japan00.htm

[13] R. Hincks, "Speech recognition for language teaching and evaluating: A study of existing commercial products", *Proceedings of ICSLP*, 733-736, 2002

[14] S. Wildner, "Learn German Now! Version 8", Software review, *CALICO Journal*, 20, 161-174, 2002.

[15] A. Mackey, J.-Y. Choi, "Review of TriplePlayPlus! English", *Language Learning & Technology* 2,19-21, 1998.

[16] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of pronunciation quality ratings and their relation to scores produced by speech recognition algorithms",*Speech Communication* 30, 109-119,2000.

[17] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality", *Speech Communication* 30, 121-130, 2000.

[18] W. Menzel, D. Herron, P. Bonaventura, R. Morton, "Automatic detection and correction of non-native English pronunciations", *Proc. of InSTIL*, Scotland, 49-56, 2000.

[19] R. Lyster, "Negotiation of Form, Recasts, and Explicit Correction in relation to error types and learner repair in immersion classrooms", *Language Learning* 48, 183-218, 1998.

[20] H. Nicholas, P.M. Lightbown, and N. Spada, "Recasts as feedback to language learners", *Language Learning* 51, 719-758, 2001.

ISBN 1-876346-48-5