ASR-based corrective feedback on pronunciation: does it really work?

Ambra Neri, Catia Cucchiarini, Helmer Strik

Centre for Language and Speech Technology (CLST)
Radboud University Nijmegen, The Netherlands
{a.neri, c.cucchiarini, h.strik}@let.ru.nl

ABSTRACT

We studied a group of immigrants who were following regular, teacher-fronted Dutch classes, and who were assigned to three groups using either a) Dutch CAPT, an ASR-based Computer Assisted Pronunciation Training (CAPT) system that provides feedback on a number of Dutch speech sounds that are problematic for L2 learners b) a CAPT system without feedback c) no CAPT system. Participants were tested before and after the training. The results show that the ASR-based feedback was effective in correcting the errors addressed in the training.

1. INTRODUCTION

The progress made in automatic speech recognition research (ASR) in the last two decades has spawned a considerable body of research into the possibilities of applying this technology to the training and testing of oral proficiency in second language learning and in particular to pronunciation, which is considered one of the most difficult skills for adults to learn in a second language. This is not surprising considering the advantages ASR-based CAPT systems can offer: extra learning time and material, specific feedback on individual errors and the possibility for self-paced practice in a private and stress-free environment. However, since it also well-known that ASR technology still has considerable limitations [1, 2] it seems legitimate to question to what extent ASR-based CAPT systems are indeed efficacious in improving pronunciation quality. To test this hypothesis we conducted a study in which we investigated whether training with an ASR-based CAPT system that provides feedback on a number of problematic speech sounds can lead to a bigger improvement of segmental quality than conventional education training.

2. ASR-BASED CAPT SYSTEM

The ASR-based CAPT system used in this study is a computer program developed at our department that provides feedback, either in Dutch or in English, on Dutch pronunciation. The system is gender-specific, because the ASR technology makes use of different parameter settings for male and female speakers. For the contents, we built on Nieuwe Buren (New Neighbours), a comprehensive CALL program used by schools for Dutch as L2 in the Netherlands and designed specifically for literate adult L2 learners with arbitrary L1s. The exercises in Dutch CAPT include role-plays, questions to be answered by uttering one of several possible sentences, and exercises requiring students to pronounce words and minimal pairs for which example pronunciations are given. The program provides feedback on eleven Dutch phonemes that appear to be problematic for speakers of different mother tongues: /ɣ/, /χ/, /ɑ/, /y/, /œy/, /aː/, /ɛi/, /ɦ/, /ʊ/, /øː/, /ɪ/ (see [3]).

Each answer provided by a student is processed by the ASR module, which first of all checks whether one of the possible answers has been spoken. In this case it immediately starts analysing it. The feedback provided consists in displaying, on the screen, the orthographic representation of the utterance pronounced by the student together with a smiley and a short comment. If the ASR algorithm finds that a phoneme has been mispronounced, the letter(s) corresponding to mispronounced phonemes are coloured red in the transcription of the utterance, a red, disappointed smiley and a message informing the student that the red sound(s) has been mispronounced are also displayed, and the students is prompted to repeat the utterance. In this way the feedback is simple and concise, and it leaves no doubt for the student that something was wrong. No more than three errors are signalled each time in order not to discourage the students. Two buttons on the interface also allow the students to listen again to their own pronunciation and to the target one, possibly focussing on the mispronounced sounds.

3. METHOD

To establish the effectiveness of our Dutch CAPT system in realistic conditions, we studied a group of immigrants who were learning Dutch in the Netherlands. The participants, who were following regular, teacher-fronted Dutch classes, were divided into three groups using either a) Dutch CAPT b) an abridged version of Nieuwe Buren, or c) no CAPT system. Participants were requested to complete questionnaires and were tested before and after the training.

To determine training effectiveness, three different types of data were used: a) the learners' appreciation of the specific CAPT received, b) expert ratings of global segmental quality, and c) expert annotations of segmental errors.

3.1 Subjects

The participants were 30 adult immigrants varying with respect to mother tongue, age, occupation and length of residence in the Netherlands who were following beginner courses of Dutch at the Radboud University Nijmegen

Language Centre (UTN). They came from 10 European, 1 Asian, and 6 African countries. They were assigned to three different groups according to instructions from the Dutch-L2 UTN coordinator, who required that students from one class would use the same computer program:

- Experimental group (EXP). Fifteen participants, 10 female and 5 male, used Dutch CAPT.
- Control group 1 (NiBu). Ten (4 female and 6 male) participants used a reduced version of Nieuwe Buren.
- Control group 2 (noXT). Five (3 female, 2 male) participants received no extra training besides the training envisaged for all UTN beginner students.

3.2 Training procedure

All three groups followed the regular classes. In addition EXP and NiBu had one extra CAPT session per week for four weeks, with each session lasting from 30 minutes to 1 hour, depending on the participant's training pace.

NiBu worked with a reduced version of Nieuwe Buren The students could record their own utterances and compare them to example utterances, but they did not receive any feedback and thus had to rely on their own auditory discrimination skills. Logfiles of each student's activities allowed the experimenter to check that all students completed all exercises as requested.

EXP used Dutch CAPT, which contained exercises that were very similar to those in Nieuwe Buren, the only difference being that it provided automatic feedback on segmental quality.

3.3 Testing procedure

3.3.1 Analysis of students' evaluations

Anonymous questionnaires were used in which participants had to indicate whether or not they agreed with a number of statements on a 5-point Likert scale and to answer two open questions. The questions concerned the accessibility of the exercises, the usability of the interface in general, the students' feelings about the usefulness of the specific CAPT for improving pronunciation, and their opinion about specific features of the system used.

3.3.2 Analysis of global segmental quality

The subjects were tested before and after the training (pre-test and post-test). To ensure that the rating process would not be influenced by possible lexical or morphosyntactical errors read speech containing every phoneme from the Dutch phonemic inventory at least once was used (phonetically rich sentences).

Six expert raters evaluated the speech independently on a 10-point scale, where 1 indicated very poor and 10 very good segmental quality. They were instructed to focus on segmental quality only, and to ignore aspects such as word stress, sentence accent, and speech rate, since these aspects were not the focus of the training the participants had received. No further instructions were given as to how to assess segmental quality. However, the raters were provided with examples of native spoken utterances and non-native spoken utterances of 'poor' segmental quality of the experiment stimuli, to help

them anchor their ratings [4]. Pre- and post-test recordings were presented in random order.

3.3.3 In-depth analysis of segmental quality

An additional, detailed analysis was carried out of the specific errors made by the participants, in order to obtain more fine-grained information on the effectiveness of the computer-generated feedback. For this investigation, auditory analyses were carried out of the participants' recordings, and annotations were made of specific segmental errors.

4. RESULTS

4.1 Analysis of students' evaluations

Overall, the responses to the questionnaires indicated a positive reaction to the two CAPT programs, with mean scores per statement ranging from a minimum of 2.4 to a maximum of 4.6 for EXP, and from 2.3 to 4.7 for NiBu. This result is in line with most studies on student appreciation of CAPT, including ASR-based CAPT [5]. More specifically, the answers indicate that the students enjoyed working with the CAPT system provided and that participants generally believed in the usefulness of the training. With respect to Dutch CAPT, eight of the 14 participants who provided comments on the system said that it was helpful, mostly in improving their pronunciation and in making them aware of specific pronunciation problems.

4.2 Analysis of global segmental quality

Before assessing the effect of the training on overall segmental quality for each group, we checked the reliability of the ratings. Inter-rater reliability was .96 and .95 for all scores and .83 and .87 when the scores assigned to the native speech fragments were removed. Intra-rater reliability was higher than .94. These coefficients are high, especially if we consider that no clear, prespecified criteria for assessment were provided.
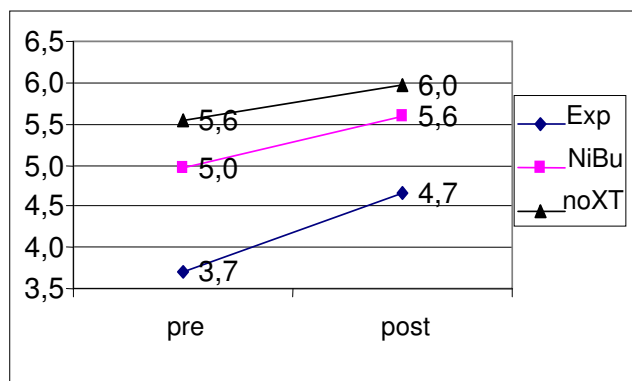

Figure 1. Average scores of global segmental quality before and after training for the three groups of participants.

We then checked whether some non-natives had received scores in the range of the natives already at pre-test. The natives were found to receive scores between 9 and 10, while

the non-native scores never fell outside the range 1-8, with a maximum average of 7.6 at pre-test.

Secondly, given the impossibility to match the treatment groups prior to the training, we examined their pre-test scores to see whether these differed significantly already before the start of the training. We carried out an analysis of variance which indicated that the group receiving no CAPT at all (noXT) had significantly higher scores than the group training with the ASR-based CAPT system (EXP).

We then looked at the average improvement made by the three groups after training, finding that overall segmental accuracy improved for all groups at post-test (see Fig. 1). Subsequently, an analysis of variance with repeated measures was conducted for the post-test condition: ANOVA_Post. The results indicated a significant effect for Test time, with $F(1, 27) = 18.806$, $p < .05$ with the post-test scores reflecting significantly greater segmental accuracy ($M = 5.19$, $SD = 1.53$) than the pre-test scores ($M = 4.42$, $SD = 1.54$). The interaction between Test time and Training group was not significant, indicating that there were no significant differences in the mean improvements of the training groups.

To summarize, these results indicate that all three groups improved overall segmental quality after the training, with the group receiving ASR-based corrective feedback showing the largest improvements, followed by the group receiving extra CAPT but no corrective feedback. However, the difference in improvements in the three groups is not statistically significant. Several explanations can be advanced for these results, e.g. the small sample size and the relatively large variation in overall segmental quality within each training group and between training groups. This variation is partly a result of the impossibility to match participants prior to the training. However, other explanations can be hypothesized for these results. For instance, it is possible that the participants did not produce errors for the phonemes addressed by the CAPT system at pre-test, in which case expecting an improvement as a result of the feedback in Dutch CAPT would be unrealistic. Another possibility is that the feedback provided was effective, but only for the 11 phonemes it targeted and that the improvement on this limited selection of phonemes did not have strong enough an impact on global segmental quality to appear in our analyses, either because the testing material did not include enough target phonemes or because the selection of target phonemes was too small.

4.3 Analysis of global segmental quality

In order to test these hypotheses, we carried out a fine-grained analysis of the segmental errors made by the participants before and after the training. An expert annotator listened to the recordings and made annotations of all segmental errors she noticed. Since making phonetic annotations is very time consuming, this task was restricted to 600 sentences.

First we checked whether the participants did indeed produce errors on the 11 target phonemes at pre-test, as we assumed when we designed Dutch CAPT. The results show that participants did mispronounce 3 to 26 (counts per participant) target phonemes at pre-test ($M = 11.23$, $SD = 5.39$), confirming the necessity of targeting at least a number of those phonemes in our automatic feedback. For EXP the range of errors per participant was 7-26 with $M = 13.93$, $SD =$

5.53; for NiBu it was 3-16 ($M = 8.1$, $SD = 4.01$); for noXT it was 4-12 ($M = 9.4$, $SD = 3.28$).

We then examined possible improvements on the 11 target phonemes and on the remaining phonemes. We checked whether and which errors decreased at post-test, and whether there were any differences between the participants who received automatic feedback and those who did not. To obtain two comparable groups differing only for 'automatic feedback', we removed noXT from these analyses.

To quantify possible decreases in errors, we calculated the percentage of errors made by each student at pre-test and post-test for each of the two types of phonemes (targeted and untargeted) relative to the amount of total phonemes of the same type in the stimuli. This examination shows that problematic errors decreased by 7.6% (absolute decrease, SD = .074) for EXP and by 1.4% (SD = .029) for NiBu. An ANOVA with repeated measures, with Training group (levels: EXP and NiBu) as between-subjects factor and Test time (levels: pre, post) as within-subjects factor, revealed a main effect for Test, $F(1, 23) = 13.319$, $p < .05$ with significantly fewer errors at post-test ($M = 11.6\%$, $SD = .056$) than at pre-test ($M = 16.8\%$, $SD = .082$). The interaction between Training and Test was also significant, $F(1, 23) = 6.175$, $p < .05$. A simple main effects analysis indicated that the factor Training had a significant effect at pre-test: $F(1, 23) = 8.18$, $p < .05$, but not at post-test. In other words, EXP was able to make a significantly faster improvement than NiBu on the targeted phonemes, catching up with NiBu.

Since it is possible that this faster improvement resulted from the fact that EXP was initially making more errors and was therefore likely to make larger improvements than NiBu [6], we also examined the errors made by both groups for the phonemes that were not targeted by Dutch-CAPT. This time a different trend appeared (see Fig. 2): While both groups produced fewer errors at post-test, the decreases in untargeted errors are much smaller and more similar across the two groups (0.7% for EXP and 1.1% for NiBu) than those for the target errors. An ANOVA with repeated measures, with Training group as between-subjects factor and Test time
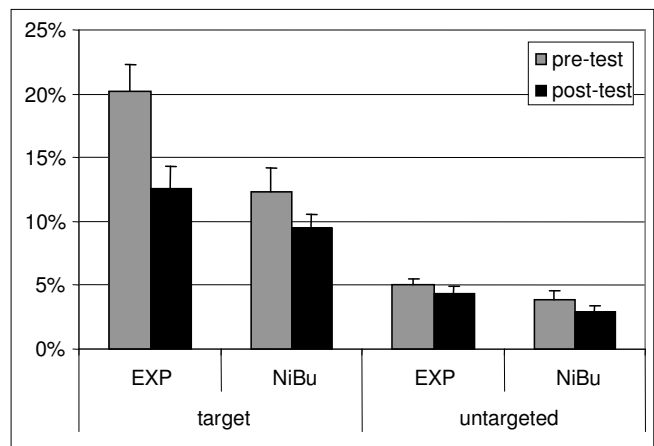

Figure 2. Percentage of pronunciation errors before and after training, for targeted and untargeted sounds.

as within-subjects factor, revealed no significant Training-Test interaction, indicating that the two groups made comparable

mean improvements on the untargeted phonemes. A main effect was found for Test, $F_{(1, 23)}$ = 10.806, p <.05, with significantly fewer errors at post-test errors (M = 3.7%, SD = .021) than at pre-test (M = 4.5%, SD = .021). No significant effect was found for Training group, confirming that, overall, the two groups made comparable proportions of untargeted errors. The mean percentages of errors on untargeted phonemes (relative to all untargeted phonemes in the stimuli) for EXP and NiBu were, respectively, 4.7% (SD = .022) and 3.4% (SD = .019).

In summary, these results show that a) the participants produced more errors for the targeted phonemes, which is an indication that these phonemes are, indeed, particularly problematic and segmental training should focus on these sounds, b) the group receiving feedback on these errors made a significantly larger improvement on the targeted phonemes, whereas no statistically significant difference was found for the phonemes for which no feedback was provided, suggesting that the automatic feedback provided in Dutch CAPT was effective in improving the quality of the targeted phonemes.

These additional analyses have evidenced specific effects that did not appear in the analysis of overall segmental quality. To understand the reasons for this discrepancy, we examined the relationship between the human ratings of global segmental quality for each participant in the groups receiving CAPT and the percentages of different errors produced by these participants at pre- and post-test. We found a strong, negative correlation between the scores assigned by the raters and the percentage of total errors per participant, $r_{(48)}$ = -.877, p <.01. This shows that the raters did indeed assess global segmental quality, i.e. segmental quality of all phonemes in the stimuli, as requested. A significant, negative correlation was also found between the scores and the percentage of untargeted errors: $r_{(48)}$ = -.863, p <.01. A significant though weaker correlation was found with the targeted errors: $r_{(48)}$ = -.645, p <.01. These results indicate that both these types of errors contributed to determining the score; however the targeted errors had less impact on the score, which is not surprising if we consider that they are less frequent (18.5%) than the untargeted phonemes (81.5%).

5. DISCUSSION AND CONCLUSIONS

The study on the effectiveness of ASR-based Dutch CAPT presented in this paper has shown that the students enjoyed using our system and that this was also efficacious in improving their pronunciation of the problematic speech sounds targeted by the automatic feedback. The fact that the effect of the corrective feedback did not appear from the global ratings of pronunciation quality, but did emerge from the fine-grained analyses of the students' utterances is a finding that deserves attention in future evaluations of such CAPT systems. Although it is undeniable that global ratings of pronunciation quality are an appropriate dependent variable, because at the end of the day CAPT should improve overall pronunciation quality, it is also clear that when evaluating systems that address specific pronunciation problems, a type of analysis with higher resolution may be required to assess the ultimate effect of the training. In our case this more detailed analysis has shown that the ASR-based feedback was

effective in improving the errors addressed in the training, but the results of the overall pronunciation ratings have made clear that this is not enough to get a significant difference in improvement with respect to the control groups. This might be due to the fact that the number of problematic sounds addressed was too small relative to the total set of sounds that may cause pronunciation errors. Recall however, that this training program was designed to be useful for students with different mother tongues. As a matter of fact the sounds addressed were those that turned to be problematic for such a miscellaneous group [3]. Possibly, these results reflect the limitations of such an approach. One can imagine that a more targeted system developed specifically for speakers with the same L1 would be more effective. Furthermore, the training period in this study was very short, perhaps too short for the learning effect to generalize to other, similar phonetic contrasts, for instance that between /o:/ and /ɔ/ or that between /e:/ and /ɛ/ as a result of training the /ɑ/-/a:/ contrast. These are issues that we intend to address in future research.

6. ACKNOWLEDGEMENTS

7. REFERENCES

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

[2] Neri, A., C. Cucchiarini, H. Strik and L. Boves. The pedagogy-technology interface in Computer Assisted Pronunciation Training, Computer Assisted Language Learning, 15:5, 441-467, 2002.

[3] Neri, A., C. Cucchiarini, and H. Strik. Segmental errors in Dutch as a second language: How to establish priorities for CAPT Proceedings of the InSTIL/ICALL Symposium, Venice, 13-16, 2004.

[4] Cucchiarini, C., Strik, H., and Boves, L. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology, Journal of the Acoustical Society of America, 107, 989-999, 2000.

[5] Mak, B.S., Ng, M., Tam, Y-C., Chan, Y-C., Chan, K-W. et al. PLASER: Pronunciation Learning via Automatic Speech Recognition, Proc. of HLT-NAACL 2003 Workshop on Building Educational Applications using Natural Language Processing, Edmonton, 23-29, 2003.

[6] Hincks, R. (2003) Speech technologies for pronunciation feedback and evaluation. ReCALL 15, 3-20, 2003.