

IMPROVING SEGMENTAL QUALITY IN L2 DUTCH BY MEANS OF COMPUTER ASSISTED PRONUNCIATION TRAINING WITH AUTOMATIC SPEECH RECOGNITION

Authors NERI, AMBRA, CUCCHIARINI, CATIA, STRIK, HELMER

ABSTRACT

Second Language (L2) researchers now agree that adult learners should aim at achieving a 'comfortably intelligible' pronunciation in order to successfully communicate in the L2 (Morley 1991), especially in the case of immigrants residing and working in the L2 country. For this purpose, it is necessary that learners receive feedback on their pronunciation from a tutor (Flege 1987). In consideration of the substantial time that pronunciation training requires from teachers, there has been a growing interest in Computer Assisted Pronunciation Training (CAPT) systems that can assess pronunciation and provide feedback automatically. CAPT systems incorporating Automatic Speech Recognition (ASR) technology are particularly attractive because of their capability to assess pronunciation quality at phoneme level.

However, little empirical evidence exists on the pedagogical effectiveness of these systems and on the contribution of different features of these systems to their overall efficacy. This is particularly regrettable for ASR-based CAPT systems because ASR technology still suffers from limitations that may result in the provision of erroneous feedback possibly leading to learning breakdowns.

In order to gain more insight into the issue of pedagogical effectiveness of ASR-based CAPT, at the Radboud University Nijmegen we designed and developed one such system and tested its effectiveness for training segmental accuracy in the pronunciation of adult learners of Dutch with different mother tongues.

The system's design derived from a rigorous study of literature and of existing systems which identified pedagogical requirements and technological possibilities (Neri et al. 2002). The resulting system contains over 100 pronunciation exercises ranging from automatic role-plays to minimal pairs, and it provides simple and easy-to-understand feedback on a selection of eleven Dutch sounds that we found to be problematic for many learners and sufficiently robust for automatic error detection (Neri et al. 2004).

To establish the system's effectiveness, we had 15 beginner learners use this system for a period of four weeks, and we compared them to two control groups: 10 students who used a similar system without automatic feedback and 5 students who received no CAPT training. All subjects also followed regular lessons. They were pre- and post-tested at approximately the same times. For each testing condition, each subject recorded two sets of phonetically rich sentences. Three different analyses were carried out. First, questionnaires were used to understand the students' impressions of the CAPT systems they used. Second, scores were obtained from six experts who rated the speech stimuli independently for overall segmental quality. Third, annotations were made of specific errors in the speech stimuli and an analysis was carried out of those errors, to discern possible differences between errors targeted by the system's feedback and other errors.

Results show that 1) students were generally satisfied with the CAPT training they

received, 2) all three groups significantly improved on overall segmental quality, 3) the group receiving ASR-based feedback made significantly larger improvements in the targeted phonemes than the two control groups. During the presentation of this work we will discuss these results and their possible pedagogical implications and suggest directions for future research.

PRESENTATION

INTRODUCTION

The progress made in ASR in the last two decades has spawned a large body of research into the possibilities of applying this technology to the training and testing of pronunciation skills in L2 learning, which are considered the most difficult skills for adults to learn in an L2. Integrating this technology within CAPT systems makes it indeed possible to offer specific feedback on individual errors, beside extra learning time and material, and self-paced practice in a private, stress-free environment. However, since it also well-known that ASR technology still has considerable limitations (Ehsani & Knodt, 1998; Neri et al., 2002) it seems legitimate to question to what extent ASR-based CAPT systems are effective in improving pronunciation quality. To investigate this issue, we compared improvements in segmental quality made by learners of Dutch who received CAPT training with ASR-based feedback with those of learners who received more conventional forms of pronunciation training.

ASR-based CAPT system

For this study, we developed an ASR-based CAPT system, Dutch-CAPT, 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. The contents are based on Nieuwe Buren (New Neighbours), a comprehensive CALL program used by schools for Dutch as L2 in the Netherlands and designed specifically for adults 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. Example pronunciations are given for all utterances. The program provides feedback on eleven Dutch phonemes that appear to be problematic for speakers of different mother tongues: /ɣ/, /χ/, /ɑ/, /y/, /œy/, /a:/, /ɛi/, /h/, /u/, /ø:/, /ɪ/ (see Neri et al. 2004).

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 by looking for the problematic phonemes. The feedback 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 corresponding letter(s) are coloured red in the transcription, a red, disappointed smiley and a message informing the student that the red sound(s) has been mispronounced are also displayed, and the student is prompted to repeat the utterance (see Figure 1). In this way the feedback is simple and concise, and it leaves no doubt that something was wrong. No more than three errors are signalled each

time in order not to discourage the students. Students can listen again to their own pronunciation and to the target one, possibly focussing on the mispronounced sounds.

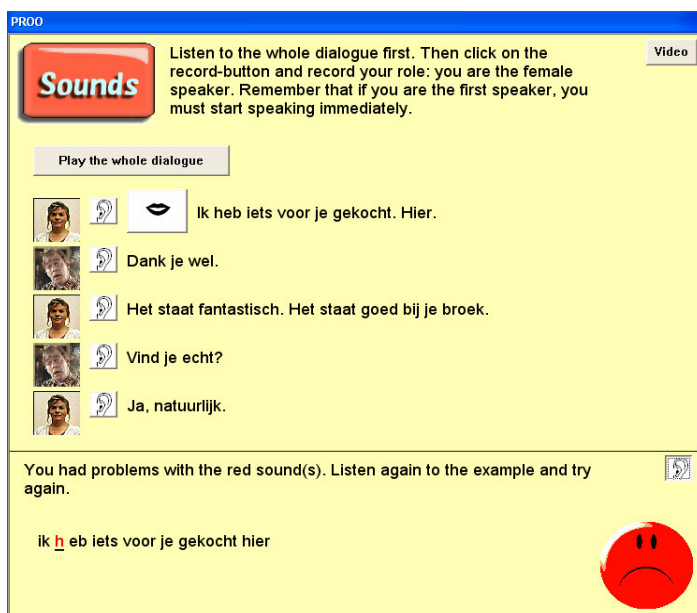


Figure 1. Snapshot of a dialogue in Dutch-CAPT in which a phoneme was mispronounced in the first utterance.

METHOD

To establish the effectiveness of Dutch-CAPT in realistic conditions, we studied a group of immigrants who were learning Dutch in the Netherlands. They were divided into three groups using either a) Dutch-CAPT b) an abridged version of Nieuwe Buren, or c) no CAPT system. Participants were tested before and after the training.

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.

Subjects

The participants were 30 adult immigrants with different L1, age, occupation, and length of residence in the Netherlands, who were following beginner courses of Dutch at the university 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 and to their availability:

- 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.

Training procedure

All three groups followed regular classes. In addition, EXP and NiBu had one extra CAPT session per week for four weeks, each lasting 30-60 minutes, depending on the participant's training pace.

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

EXP used Dutch-CAPT, with exercises that were comparable to those in Nieuwe Buren, the only difference being the automatic feedback on segmental quality provided by Dutch-CAPT.

Testing procedure

Analysis of students' evaluations

Students were asked to complete anonymous questionnaires by indicating whether or not they agreed with statements on a 5-point Likert scale and by answering two open-ended 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.

Analysis of global segmental quality

The subjects were tested before (pre-test) and after the training (post-test). To ensure that the rating process would not be influenced by lexical or morphosyntactical errors, read speech containing all Dutch phonemes 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. To help them anchor their ratings (Cucchiari et al. 2000), the raters were provided with examples of native spoken utterances and non-native spoken utterances of 'poor' segmental quality of the experiment stimuli. Pre- and post-test recordings were presented in random order.

In-depth analysis of segmental quality

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

RESULTS

Analysis of students' evaluations

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 congruent with other studies on ASR-based CAPT (Mak et al. 2003). The answers show that the students enjoyed working with the CAPT system provided and that they generally believed in its usefulness. 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.

Analysis of global segmental quality

First of all, 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.

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 of 7.6 at pre-test.

Secondly, given the impossibility of matching the groups before the training, we examined their pre-test scores to see whether these differed significantly already prior to the training. An ANOVA with post-hoc comparisons 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 examined the average improvement made by the groups after training, finding that overall segmental accuracy improved for all groups at post-test (see Figure 2). Subsequently, an ANOVA with repeated measures was conducted indicating 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 Test time x Training group interaction was not significant, indicating that there were no significant differences in the mean improvements of the training groups.

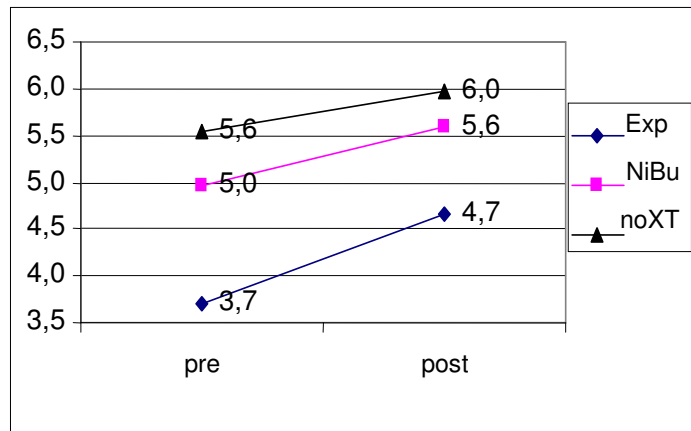


Figure 2. Average scores of global segmental quality before and after training for the three groups of participants.

To summarize, all three groups improved overall segmental quality after the training, with Exp showing the largest improvements, followed by NiBu. However, the difference in improvements in the three groups is statistically nonsignificant. Several explanations can be advanced for these results, e.g. the small sample size and the relatively large within-group and between-group variation in segmental quality. This variation partly results from the impossibility of matching participants before the training. However, it is also possible that the participants did not produce errors for the phonemes addressed by the CAPT system already 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 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.

In-depth analysis of segmental quality

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.

We subsequently checked whether the participants did indeed produce errors on the 11 targeted 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) targeted phonemes at pre-test ($M=11.23$, $SD=5.39$), confirming the necessity of targeting at least a number of those phonemes. 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 all 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 total phonemes of the same type in the stimuli. Problematic errors seem to have decreased by 7.6% ($SD=.074$) 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 this faster improvement could have resulted from the fact that EXP was initially

making more errors and was therefore likely to make larger improvements than NiBu (Hincks, 2003), 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 Figure 3): 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 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 ($M=3.7\%$, $SD=.021$) 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% ($SD=.019$).

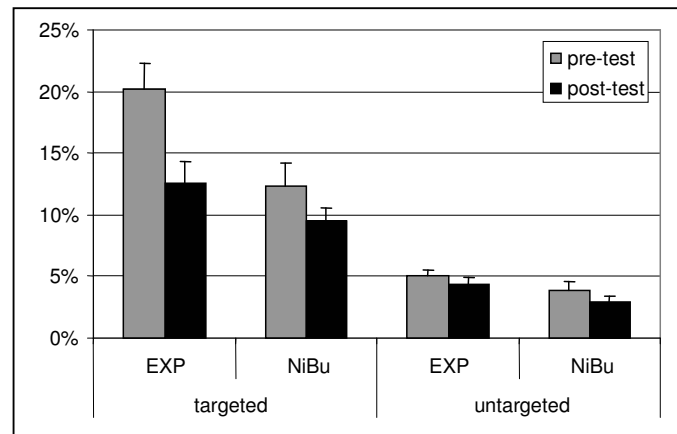


Figure 3. Mean percentages of pronunciation errors before and after training, for targeted and untargeted sounds.

These results show that a) more errors were produced for the targeted phonemes, which confirms that these phonemes are particularly problematic and segmental training should focus on these sounds, b) the group receiving feedback on these errors made a significantly larger improvement on these phonemes, whereas no statistically significant difference was found for the phonemes for which no feedback was provided, suggesting that the 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 raters' scores and the percentage of total errors per participant, $r = -.877$, $p < .01$. Thus 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: $r = -.863$, $p < .01$. A significant though weaker correlation was found with the targeted errors: $r = -.645$, $p < .01$. These results indicate that both types of errors contributed to determining the score; however the targeted errors had less impact on it, which is not surprising if we consider that the targeted *phonemes* are less frequent (18.5%) than the untargeted phonemes (81.5%).

DISCUSSION AND CONCLUSIONS

This study has shown that the students enjoyed using our system and that this was also efficacious in improving their pronunciation of the problematic phonemes targeted by the automatic feedback. The fact that the effect of the 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 CAPT systems. Although it is undeniable that global ratings of pronunciation quality are an appropriate measure, because CAPT should ultimately improve overall pronunciation quality, it is also clear that when evaluating systems addressing specific pronunciation problems, an 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 shown that this is not enough to obtain 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 Dutch-CAPT was designed to be useful for students with different L1s, and the phonemes addressed here were those that appeared to be problematic for such a miscellaneous group (Neri et al. 2004). Possibly, these results reflect the limitations of such an approach. A more targeted system tailored to speakers with the same L1 might be more effective. Furthermore, the training intensity in this study was very low, perhaps too low for the learning effect to generalize to other, similar phonetic contrasts in Dutch, for instance with respect to vowel length. These are issues that we intend to address in future research.

ACKNOWLEDGEMENTS

This study was supported by the Dutch Organization for Scientific Research (NWO). We are indebted to Ming-Yi Tsai, Febe de Wet, Michela Hulbosch, Louis ten Bosch, Christophe van Bael, Joop Kerkhoff, and Albert Russel for their help building Dutch-CAPT and to Loe Boves and Toni Rietveld for their comments on this paper. Many thanks also go to the students and teachers at UTN.

REFERENCES

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

Ehsani, F. & Knodt, E. (2002) Speech technology in computer-aided learning: Strengths and limitations of a new CALL paradigm. *Language Learning and Technology*, 2, 45-60.

Flege, J.E. (1987). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics*, 15, 47-65.

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

Mak, B.S., Ng, M., Tam, Y-C., Chan, Y-C., Chan, K-W. et al. (2003) PLASER: Pronunciation Learning via Automatic Speech Recognition. In *Proceedings of HLT-NAACL 2003 Workshop*, Edmonton, 23-29.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly*, 25, 481-519.

Neri, A., Cucchiari, C. & Strik, H. (2004). Segmental errors in Dutch as a second language: How to establish priorities for CAPT. In *Proceedings of the INSTIL/ICALL Symposium*, Venice, 13-16.

Neri, A., Cucchiari, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in Computer Assisted Pronunciation Training. *Computer Assisted Language Learning*, 15, 441-467.

BIODATA

Ambra Neri has worked as a PhD student on a project aimed at establishing the usefulness of Automatic Speech Recognition (ASR) for training pronunciation in Dutch (L2). Her work has focussed in particular on automatic feedback on segmental quality. For her project, A. Neri has been supervised by co-authors Catia Cucchiari and Helmer Strik, who have worked extensively on automatic assessment of oral proficiency of non-native speakers by means of ASR technology. Other research areas in which Dr. Cucchiari and Dr. Strik have worked include pronunciation variation, ASR for Dutch, ASR and dysarthric speech.

CONTACT

Ambra Neri

CLST, Radboud University Nijmegen,
Erasmusplein 1
6500 HT Nijmegen
The Netherlands

A.Neri@let.ru.nl
Http://lands.let.ru.nl