The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch

AMBRA NERI, CATIA CUCCHIARINI AND HELMER STRIK
*Centre for Speech and Language Technology, Department of Linguistics,
Radboud University Nijmegen, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands*
(*email: C.Cucchiarini@let.ru.nl; ambra.neri@gmail.com; strik@let.ru.nl*)

Abstract

Although the success of automatic speech recognition (ASR)-based Computer Assisted Pronunciation Training (CAPT) systems is increasing, little is known about the pedagogical effectiveness of these systems. This is particularly regrettable because ASR technology still suffers from limitations that may result in the provision of erroneous feedback, possibly leading to learning breakdowns. To study the effectiveness of ASR-based feedback for improving pronunciation, we developed and tested a CAPT system providing automatic feedback on Dutch phonemes that are problematic for adult learners of Dutch. Thirty immigrants who were studying Dutch were assigned to three groups using either the ASR-based CAPT system with automatic feedback, a CAPT system without feedback, or no CAPT system. Pronunciation quality was assessed for each participant before and after the training by human experts who evaluated overall segmental quality and the quality of the phonemes addressed in the training. The participants' impressions of the CAPT system used were also studied through anonymous questionnaires. The results on global segmental quality show that the group receiving ASR-based feedback made the largest mean improvement, but the groups' mean improvements did not differ significantly. The group receiving ASR-based feedback showed a significantly larger improvement than the no-feedback group in the segmental quality of the problematic phonemes targeted.

Keywords: pronunciation training; automatic speech recognition; pedagogical effectiveness

1 Introduction

The last decade has witnessed a growing popularity of Computer Assisted Pronunciation Training by means of Automatic Speech Recognition because these systems can provide automatic and individualized feedback in a private environment. However, little research has been carried out on the effectiveness of ASR-based feedback in improving a learner's pronunciation. This is a common problem in Computer Assisted Language

Learning (CALL) research in general (Felix, 2005), and it is possibly related to the practical difficulties in recruiting participants who want to train regularly and over an extended period of time with a system whose usefulness still has to be ascertained. The need for such information is acute in the case of ASR-based CAPT because state-of-the-art ASR technology is known to suffer from limitations that can result in the occasional provision of erroneous feedback to the learner, possibly compromising the learning process and outcome. This is all the more serious for ASR-based CAPT providing automatic evaluation of the quality of individual segments, which is one of the most challenging tasks for this technology (Kim, Franco & Neumeyer, 1997).

The purpose of this study is to examine the pedagogical effectiveness of ASR-based CAPT for adult, beginner learners of Dutch with different mother tongues. More precisely, we aim at establishing whether training with a CAPT system providing simple and easy-to-understand ASR-based feedback on a selection of problematic phonemes can improve segmental quality, i.e. the pronunciation, of those phonemes and whether the possible improvement can indeed be ascribed to the specific automatic feedback provided.

This paper is organized as follows: first, pedagogical requirements for technologically reliable ASR-based CAPT are briefly sketched; the specific ASR-based CAPT system developed for this study is presented; the experiment carried out is subsequently described, followed by results and considerations on methodological issues in the measurement of the effectiveness of ASR-based feedback for pronunciation training.

2 Studies on the effectiveness of ASR-based CAPT

The few studies available on the effectiveness of ASR-based CAPT providing automatic feedback at segmental level differ substantially in the type of CAPT system and in the experimental design, making it difficult to draw an uniquevocal conclusion from their results.

Akahane-Yamada, Dermott, Adachi, Kawahara, and Pruitt (1998) examined the effectiveness of two different types of automatic feedback in helping Japanese learners improve perception and production of the 'l/r' contrast in English. One, administered to ten participants for three days, consisted of spectrographic representations of the trainees' speech and of the model-talker's speech. The other form of automatic feedback, which was provided to two participants for two hours, consisted of scores obtained with an ASR system. The stimulus material contained a list of minimal pairs to be read aloud. Native speakers of American English auditorily evaluated the intelligibility of the consonants by means of a two-alternative, forced choice task and rated the 'goodness of the productions' on a 7-point scale. The results indicated improvements both in intelligibility and goodness of production, and a small increase in perception for both groups of participants.

Mayfield Tomokyio, Wang and Eskenazi (2000) examined the improvement made by sixteen university students with different native languages in the pronunciation of the English voiced and voiceless fricatives /Δ/ and /T/. All students practised for two hours during two to three weeks in an immersion situation. Eight students used *Fluency*, an ASR-based system which identifies errors automatically and offers suggestions for correctly pronouncing the phonemes targeted. The control group (n=8) received the

same type of training by a teacher. A noticeable error reduction was found for the fricatives in different phonetic contexts for both groups. Based on these results, the authors conclude that the type of training offered by *Fluency* was effective and that it could be successfully applied to other phonemes too.

Hincks (2003, 2005) studied the effectiveness of *Talk To Me-English* (henceforth TTM-E) and compared nine students who used *TTM-E* on average 12.5 hours with eleven students who had only followed the traditional course in the previous semester. To measure pronunciation quality before and after training, Hincks used an ASR-based automatic telephone test lasting ten minutes. Neither group of participants showed significant mean improvements in global pronunciation quality after ten weeks of training. Analyses of individual scores revealed that the participants with an initial strong foreign accent improved much more in the experimental group than in the control group, indicating that the less proficient students benefited from practising with *TTM-E*. By contrast, more proficient participants in the experimental group got worse. In a later analysis, Hincks (2005) compared the results of the automatic test with those given by a pool of raters and found that the results coincided for only eleven of the twenty-four participants. However, the trend according to which the students with a more intrusive accent improved and the more proficient ones got worse also emerged from the human assessment.

A recent study was carried out on *PLASER*, an ASR-based CAPT system for Cantonese Chinese learners of English (Mak, Siu, Ng, Tam, Chan, Y-C., Chan, K-W. *et al*., 2003) focussing on the pronunciation of confusable phonemes. Pre-test and post-test recordings of sixty words were available for 210 participants. Although they do not specify how it was measured, the authors report a significant mean improvement in pronunciation accuracy, with 73% students improving on average by 4.53% (absolute increase) and 27% getting worse on average by 2.68% for unknown reasons. The majority of the students appeared to appreciate this type of training and to prefer it to traditional pronunciation classes; the teachers' impressions were also very positive.

Although these studies mostly indicate a positive effect of ASR-based CAPT, it is often unclear how the improvement was measured and whether the improvement can be attributed to the automatic feedback provided. In the remainder of this paper we will address and discuss most of these issues in some detail.

3 The ASR-based CAPT system and the corrective feedback provided

For this experiment, *Dutch-CAPT,* an ASR-based CAPT system, was built on the basis of studies available on L2 learning and on CAPT, as well as of research we carried out for the specific purpose of providing CAPT to learners of Dutch.

The analysis in Neri, Cucchiarini, Strik and Boves (2002) indicated that some ASR-based CAPT programs provide feedback that is not always clear. For instance, they show waveforms, which look impressive, but are incomprehensible to most students. Therefore, in Dutch CAPT feedback was designed so as to provide an overt and clear indication that an error had occurred.

In addition, we opted for achieving intelligibility rather than accent-free pronunciation. Furthermore, we focussed on segmental feedback because one of the main advantages of Automatic Speech Recognition is its possibility to segment an

utterance in phonemes so that feedback can be provided at phoneme level; however, as this assessment is based on relatively few measurement points, this task is a challenging one for ASR-based CAPT technology (Kim *et al.,* 1997).

Since Dutch L2 instruction in the Netherlands is mainly administered in classes with speakers of different mother tongues, we decided to focus on a mixed group of learners. To increase training efficiency, we designed the following criteria to select the segmental errors to address in the training: the error should be (1) perceptually salient, (2) frequent, (3) common across speakers of various mother tongues, (4) persistent over time, (5) potentially hampering to communication, and (6) suitable for robust automatic detection. These criteria were applied in analysing three different speech databases of read and extemporaneous speech in Dutch as second and foreign language. The 116 speakers were beginner, intermediate, and proficient learners with different mother tongues. Preliminary tests based on these criteria yielded a selection of eleven target Dutch phonemes: /Ψ/, /ξ/, /A/, /ψ/, /⌐ψ/, /α⌐/, /Eʊ/, /η/, /ʊ/, /O⌐/, /I/ (see Neri *et al.*, 2006).

The didactic content was adapted from *Nieuwe Buren* (New Neighbours) 2.1, a comprehensive CALL system widely used for Dutch as L2 for literate adult learners with different L1s, which contains videos of a tailor-made, sitcom-like story presenting real-life situations.

The content of Dutch-CAPT was divided into four units, each containing one video from *Nieuwe Buren*, and approximately 25 exercises based on the video which were to be completed in a sequential, constrained order to ensure that each student received a similar amount of training. These exercises include a total of 22 role-plays (see Figure 1), 46 aural or written questions to be answered orally by recording one of several possible answers, and 38 exercises requiring the student to pronounce specific words for which example pronunciations are given, including minimal pairs (see Figure 2).

The students' answers are processed by a gender-specific ASR module that first of all checks whether one of the possible answers has been spoken, in which case it immediately starts analysing pronunciation quality. An algorithm based on the Goodness of Pronunciation method (Witt & Young, 2000), which makes use of binary thresholds, determines whether a phoneme was correctly pronounced or not. The feedback provided consists in displaying, on the screen, the orthographical transcription of the utterance pronounced by the learner, together with a 'smiley' and a short written comment (see Figure 2).

If a phoneme has been mispronounced, the letter(s) corresponding to that phoneme is (are) coloured red and underlined in the transcription, which is accompanied by a red, disappointed smiley and by a message informing the student of the error and prompting him/her to repeat the utterance (see Figure 1). In this way the feedback is simple and concise, and it can be unmistakeably perceived by the student as corrective. No more than three errors are signalled in an utterance, in order not to discourage the students. Two buttons on the graphical user interface allow the students to listen again to their own pronunciation and to the target one (see Figure 2). This type of feedback is meant to make the learner focus on the erroneous phonemes and cater for perceptual as well as productive training: the provision of feedback at every successive attempt should help the learner gradually adjust his/her pronunciation towards the target one. For a functional description of Dutch CAPT and the algorithm's scoring accuracy, see Neri (2007).

Fig. 1. Snapshot of a dialogue in Dutch-CAPT in which a phoneme was mispronounced in the first utterance. The English translation of the dialogue is as follows: -I have bought you something. Here. -Thank you. -It suits you perfectly. It matches your trousers. –Do you really think so? –Yes, of course.

4 Method

To establish the effectiveness of ASR-based CAPT in realistic conditions, we studied a group of immigrants who were learning Dutch in the Netherlands. Three different types of analyses were conducted: (a) of global segmental quality based on human ratings, (b) of the specific phonemic errors based on (manual) annotations, and (c) of the learners' appreciation of the specific CAPT received.

4.1 Participants

The participants were thirty adult immigrants with varying mother tongue (see Table 1), age, reason for studying Dutch, and length of residence in the Netherlands. They were all following beginner courses of Dutch at UTN, the language centre of the Radboud University Nijmegen. They were non-randomly assigned to three different groups according to instructions from the Dutch-L2 coordinator at UTN, who required that students from one class would use the same computer program:
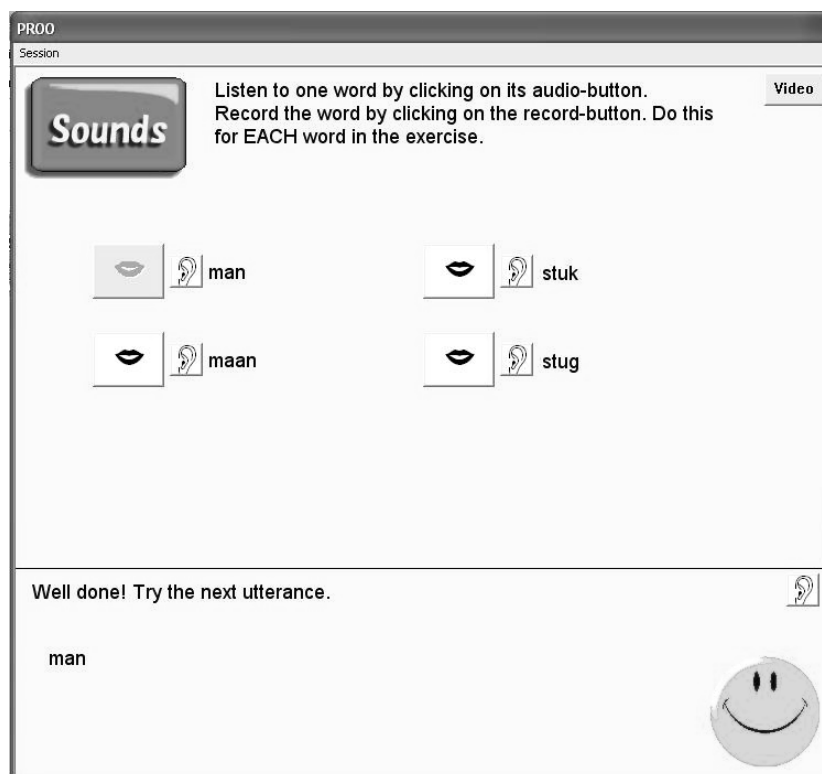
Fig. 2. Snapshot of a minimal-pair exercise in Dutch-CAPT. The English translation of the words is as follows (clockwise): piece, stiff, moon, man.

(a) the experimental group using *Dutch-CAPT (EXP)* consisted of fifteen (ten female and five male) participants, with mean age 29 ($SD = 4.6$), and length of residence in the Netherlands ranging from 2 months to 5.5 years ($M = 16.7$ months, $SD = 19.1$);

(b) the control group using a reduced version of *Nieuwe Buren (CTRL_NB)* consisted of ten (four female and six male) participants, with mean age 32 ($SD = 8.9$) and length of residence between 3 months and 4.9 years ($M = 10.2$ months, $SD = 17.2$);

(c) the control group receiving no extra training *(CTRL_NO)* consisted of three female and two male participants with mean age 23 ($SD = 2.6$) and length of residence ranging from 1 month to 1.7 years ($M = 8.4$ months, $SD = 7.8$).

All participants had obtained a diploma or a university degree. They had been assigned to a beginner course by an experienced UTN teacher who interviewed them. The courses encompassed four to six hours per week of collective classes with a Dutch teacher, supplemented by self-study sessions in the language lab. During these sessions, participants could complete pencil-and-paper exercises and work with *TaalRecorder* (an application with a simple, tape recorder-like computer interface providing audio

Table 1 *Distribution of participants per training group and mother tongue (L1)*

L1	Training group			Total
	EXP	**CTRL_NB**	**CTRL_NO**	
Arabic	6			6
Bengali		1		1
Catalan		2		2
English	1	1	1	3
German		1		1
Greek		2		2
Hebrew	1			1
Italian		1	1	2
Lithuanian			1	1
Polish	2	1	2	5
Russian	1			1
Spanish	1			1
Swedish	1			1
Turkish	2			2
Ukrainian		1		1
Total	15	10	5	30

samples of Dutch sentences and a record- and playback-feature). By the time they were pre-tested, most of the participants had followed Dutch classes at UTN for four weeks. All participants received monetary compensation for their participation in the experiments.

4.2 Training procedure

All three groups followed the regular classes. CTRL_NB and EXP had one CAPT session per week for four weeks, with each session lasting 30 minutes to one hour, depending on the participant's training pace. The sessions were individual and took place in the language lab under the supervision of the experimenter (the first author). CTRL_NB worked with a reduced version of *Nieuwe Buren* that matched *Dutch-CAPT*. These 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 allowed the experimenter to check that all students completed all exercises as requested. EXP used *Dutch-CAPT*, which only differed from Nieuwe Buren for the provision of automatic feedback on segmental quality.

4.3 Testing procedure

4.3.1 Analysis of global segmental quality

The participants were tested before and after the training (pre-test and post-test). They read aloud from a computer screen two sets of phonetically rich sentences, i.e. sentences containing every phoneme of Dutch at least once. One set (henceforth 'simple stimuli')

contained five short sentences with relatively short and frequent words with no difficult consonant clusters. The other set (henceforth 'complex stimuli') contained five longer sentences, longer and less familiar words and consonant clusters. These two sets were used to study possible effects of the stimuli's characteristics on the potential improvement in segmental quality. Read speech was chosen to ensure that the rating process would not be influenced by other factors, such as lexical or morphosyntactical errors that might emerge in a spontaneous speech production task. The recordings took place in the language lab by means of head-mounted microphones, and the speech was sampled at 16kHz.

Six Dutch-L2 expert raters evaluated, independently and in random order, the utterances on a 10-point scale, where 1 indicated very poor and 10 very good segmental quality. The raters, two teachers and four linguists who had worked extensively on L2-Dutch, had never had any contact with the participants. They were instructed to focus on segmental quality only, ignoring aspects such as word stress, sentence accent, and speech rate, which were not the focus of the training. No further instructions were given as to how to assess segmental quality. However, examples were provided of native and non-native spoken utterances of opposite segmental quality, to help the experts anchor their ratings (Cucchiarini, Strik & Boves, 2000).

In order to limit the size of the evaluation task, the speech material was divided into three parts and each part was assigned to two raters, in such a way that the speakers to be evaluated by each pair of raters would vary as much as possible with respect to pronunciation quality, mother tongue, and gender. To calculate inter-rater reliability, we followed the procedure adopted in Cucchiarini *et al*. (2000) by assigning a portion of the speech material to all raters: one female and one male native speaker (NS) of Dutch, who served as the upper bound for the ratings, and five non-native speakers (NNS) were thus evaluated by all raters. As a result, each rater was eventually assigned the two NS, the five common NNS, and nine unique NNS.

4.3.2 In-depth analyses of segmental quality

In-depth analyses of the specific errors made by the participants were carried out to obtain more fine-grained information on the effectiveness of the computer-generated feedback.

4.3.3 Analysis of students' evaluations

The participants in EXP and CTRL_NB also expressed their opinions about their training programs through anonymous questionnaires in which they indicated their agreement with a number of statements on a 1-5 Likert scale and answered two open-ended questions.

5 Results and discussion

5.1 Analysis of global segmental quality

5.1.1 Reliability of ratings

Following the procedure described in Cucchiarini *et al*. (2000), we checked the reliability of the ratings. Inter-rater reliability was .96 and .95 for all scores and .83 and

Table 2 *Reliability coefficients (Cronbach's alpha) for the two raters groups.*

	Intra-rater			Inter-rater	
	Rater 1	**Rater 2**	**Rater 3**	**NNS & NS**	**NNS only**
Group A	.98	.94	1.00	.96	.83
Group B	.99	.98	.98	.95	.87

.87 when the scores assigned to the native speech fragments were removed. Intra-rater reliability was higher than .94 (see Table 1). These coefficients are high, especially if we consider that no clear, prespecified criteria for assessment were provided. Consequently, we decided to combine all scores into one score for each set of sentences produced by each participant (see Neri, 2007).

5.1.2 Global segmental quality

At pre-test, the NS were found to receive scores ranging between 9 and 10, while the NNS scores never fell outside the range 1-8, with a maximum of 7.6 at pre-test (*M* = 4.42, *SD* = 1.56 and M = 4.43, *SD* = 1.68 for the simple and complex stimuli, respectively).

Given the impossibility of matching 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 (ANOVA with repeated measures) with Training group as the between-subjects factor (levels: EXP, CTRL_NB, CTRL_NO) and Stimulus type (levels: simple, complex) as within-subjects factor. The analysis yielded a significant main effect of Training group with $F(2, 27)$ = 4.498, p <.05 (Partial η^2 = .250). The mean scores (with standard deviations in brackets) for EXP, CTRL_NB and CTRL_NO are 3.69 (1.42), 4.96 (1.33), and 5.55 (1.34), respectively. No significant effect was found for Stimulus type, nor was the interaction between Training group and Stimulus type significant, indicating that the complexity of the stimuli did not affect the overall differences between the three training groups. Post-hoc comparisons (Tukey's HSD test) of the groups' overall scores indicated a significant difference only between EXP and CTRL_NO (p <.05), i.e. between the group training with the ASR-based CAPT system and the group receiving no CAPT at all.

We then examined the data to discern global trends. First, we looked at the average improvement made by the three groups at post-test, finding that overall segmental accuracy improved for all groups (see Figure 3). The largest improvement was made by EXP, i.e. the group using the ASR-based CAPT system with feedback, followed by CTRL_NB. However, as can be seen from the SD bars in Figure 3, EXP also presents the largest within-group variation in all testing conditions.

We also examined the participants' performance with respect to the specific types of stimuli and found that all groups improved for both the simple and the complex stimuli. However, no common trend can be evidenced with respect to stimulus type: the groups
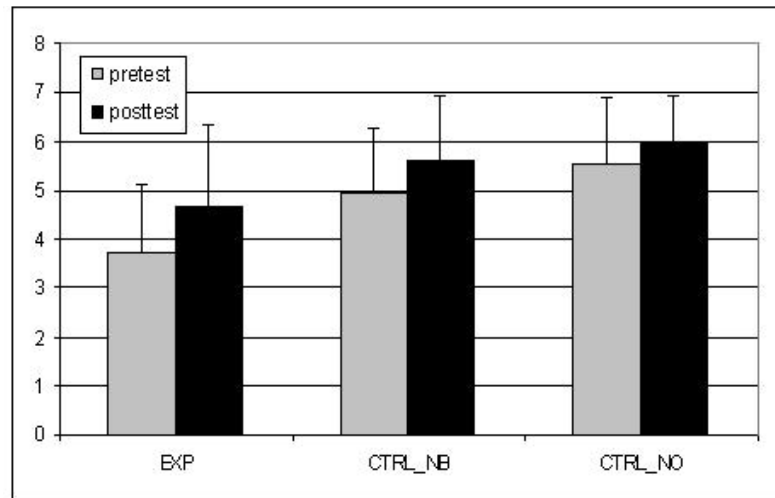
Fig. 3. Mean scores (based on a 10-point scale) for each training group at each testing condition. The means were computed by first averaging the individual scores for simple and complex stimuli. Bars indicate one SD from the mean.

seem to differ in the way they perform on the different stimuli.

The participants' scores were then submitted to an ANOVA with repeated measures. The within-subjects factors were Test time (levels: pre-test, post-test) and Stimulus type (levels: simple, complex), the between-subjects factor was Training group (levels: EXP, CTRL_NB, CTRL_NO). The results indicate 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. The mean scores for EXP, CTRL_NB and CTRL_NO were 4.18 ($SD = 1.49$), 5.28 ($SD = 1.29$), and 5.76 ($SD = 1.11$), respectively, in line with the pre-test data. Subsequent post-hoc comparisons (Tukey's HSD test) with an alpha level of .05 did not show any significant difference between the three groups. However, the main effect found for Training group, the groups' mean scores, and the significant difference between EXP and CTRL_NO at pre-test suggest that these two groups are still rather different in segmental quality at post-test. No significant effect was found for Stimulus type. This result, together with the lack of significant interactions involving Stimulus type in this analysis, confirms that the different types of stimuli did not affect segmental quality in a significant way across training groups or testing condition, or within training groups at different testing conditions.

To summarize, these results indicate that the mean improvements made by the three groups on global segmental quality after the training were not significantly different from each other and that the complexity of the stimuli did not play a role. Several explanations can be hypothesized for the lack of a significant difference between the mean improvements of the three groups. First of all, the small sample size and the relatively large variation in overall segmental quality within each training group and
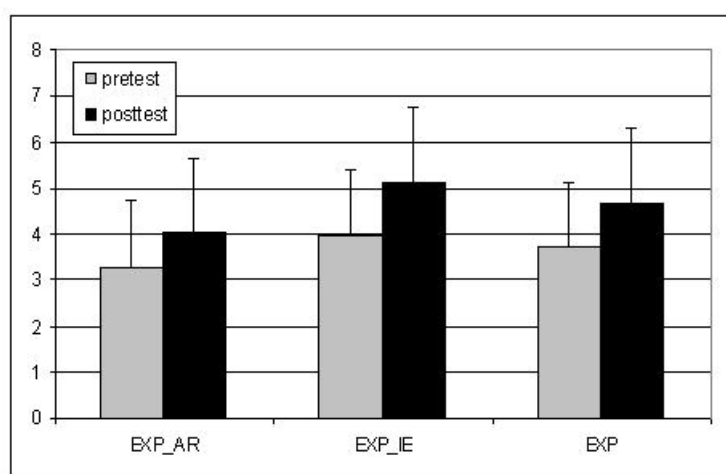
Fig. 4. Group mean scores (based on a 10-point scale) at each testing condition for the speakers of Arabic in EXP (EXP_AR), for the speakers of Indo-European languages in EXP (EXP_IE), and overall (EXP). The bars indicate one SD from the mean.

between training groups (see Figure 3). This variation is partly a result of the impossibility of matching participants prior to the training, to comply with the instructions from the UTN coordinator. This problem is reflected in the uneven distribution of mother tongues across the training groups: EXP included six native speakers of Arabic, whereas CTRL_NB and CTRL_NO included only speakers with Indo-European languages as mother tongues. Previous research (Bongaerts, 2001; Purcell & Suter, 1980) has suggested that speakers of languages that are typologically distant from the target language will find it more difficult to learn that language and will thus be slower in their progress. This hypothesis might explain the data in Figure 4: the speakers of Arabic achieve lower mean scores and make smaller improvements than their Indo-European counterparts in each testing condition.

The lack of a statistically significant advantage for the students receiving automatic feedback and extra CAPT in general, with respect to the students in CTRL_NO, might indicate that the CAPT provided was simply not intensive enough, regardless of the participants' mother tongue. The number and the frequency of the extra training sessions (recall that these were, in total, four weekly sessions of 30 to 60 minutes each) were indeed low compared to the number and frequency of teacher-led training sessions (four to six hours per week). More intensive training is more likely to impact on overall pronunciation quality (see also Macdonald, Yule & Powers, 1994; Precoda, Halverson & Franco, 2000) and to generalize to other phonemes. Other possible explanations might be that the targeted phonemes were not problematic for the participants, in which case expecting an improvement would be unrealistic, or that the automatic feedback was not effective in improving those errors, or that the feedback was effective for the eleven phonemes, but the number of targeted phonemes was too small to impact on overall segmental quality. In order to investigate these issues, more detailed analyses were carried out.

5.2 *In-depth analyses of segmental quality*

An expert annotator listened to the recordings and indicated whenever a phoneme (either targeted or untargeted) was pronounced incorrectly. Errors on targeted and untargeted phonemes for each participant at pre-test and post-test were thus tallied.

The results show that participants did mispronounce 3 to 26 targeted phonemes at pre-test ($M = 11.23$, $SD = 5.39$) out of a total of 69 targeted phonemes in the pre-test stimuli. For EXP the range of errors on targeted phonemes per participant was 7-26, ($M = 13.93$, $SD = 5.53$); for CTRL_NB it was 3-16 ($M = 8.1$, $SD = 4.01$); for CTRL_NO it was 4-12 ($M = 9.4$, $SD = 3.28$). These data confirm that the targeted phonemes were problematic at pre-test, thus the selection of targeted phonemes was correct.

We then examined possible improvements on the targeted and untargeted phonemes in the participants who received automatic feedback and those who did not. In order to have two comparable groups that would only differ in respect of the presence of the variable 'automatic feedback' in the training, we compared EXP with CTRL_NB. We then 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. For instance, if the total number of phonemes in the stimulus material was 372, of which 69 were targeted phonemes, and a student produced a total of 21 errors on targeted phonemes at pre-test, the student's percentage of targeted errors at pre-test would be $(21/69) \times 100\%$, in this case 30.4%. This was necessary because the occurrences of targeted and untargeted phonemes in the stimuli greatly differed: the simple and complex stimuli contained 69 targeted phonemes and 303 untargeted phonemes.

Comparisons between the percentages of total errors per participant at pre-test and post-test with the human ratings of global segmental quality for each participant in the two groups show a strong, negative correlation, $r(48) = -.877$, $p < .01$, indicating that the annotations of all errors and the ratings of global segmental quality were in agreement. The examination of relative error percentages shows that problematic errors decreased by 7.6% (mean calculated over absolute individual decreases, $SD = .074$) for EXP and by 1.4% ($SD = .029$) for CTRL_NB. We submitted the percentages of targeted and untargeted errors at pre-test and post-test for EXP and CTRL_NB to an ANOVA with repeated measures. The factors were Training group as between-subjects factor and Test time as within-subjects factor. Box's M Test for the ANOVA on targeted errors was significant, with $F(3, 15.679) = 4.701$, $p = .003$; thus, the homogeneity of covariances assumption was violated.

Therefore, non-parametric tests were performed for these analyses. We examined the overall improvement of all the subjects on targeted and untargeted errors by carrying out a Wilcoxon Signed Ranks test on the subjects' scores for targeted and untargeted errors. Outliers from the EXP group were removed before performing the tests and the alpha value was set at .025. The results indicate an overall improvement on both types of errors at post-test: for targeted errors, the test yielded a z of –3.150, $p = .001$ (one-tailed); for untargeted errors, the test yielded a a z of –3.360, $p = .000$ (one-tailed).

We subsequently examined the difference between the improvements of each group on the two types of errors by using the Wilcoxon Rank-sum test. Outliers were removed before carrying out the analysis and the significance level was adjusted to .0125. The test on targeted errors indicated a significant difference between EXP and CTRL_NB
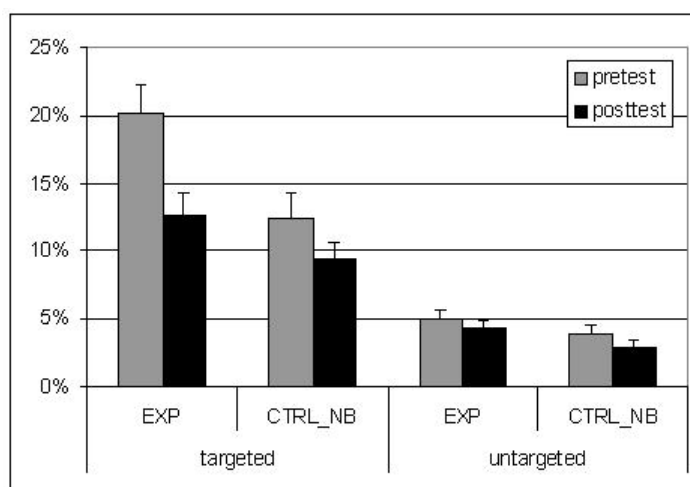
Fig. 5. Mean error percentages (and SEMs) for errors on the targeted and untargeted phonemes. The percentages represent the proportion of the occurrences of a certain error type with respect to the total number of phonemes of the same type in the stimuli.

(z=-2.827, p = .002, one-tailed), with EXP making a significantly larger improvement than CTRL_NB on segmental quality of the targeted phonemes. This faster improvement might be a consequence of the fact that EXP was initially making more errors and therefore had more room for improvement than CTRL_NB, as found in De Bot (1983) and Hincks (2003, 2005). Therefore, we also examined the errors made by both groups for the phonemes that were not targeted by the feedback in Dutch-CAPT. This time a different trend appeared (see four rightmost bars in Figure 5): 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 CTRL_NB) than those for the targeted errors, and in this case the largest decrease is made by CTRL_NB, not by EXP.

The Wilcoxon Rank-sum test on untargeted errors revealed no significant difference between the sum of ranks of the two groups, indicating that the two groups made comparable mean improvements on untargeted phonemes. The mean percentages of errors on untargeted phonemes (relative to all untargeted phonemes in the stimuli) for EXP and CTRL_NB were, respectively, 4.7% (*SD* = .022) and 3.4% (*SD* = .019).

In summary, these results indicate that (a) the participants produced relatively more errors for the targeted phonemes, suggesting that these phonemes are particularly problematic and that segmental training should focus on these sounds, (b) the group receiving feedback on these errors made a significantly larger improvement on targeted phonemes than the group receiving no automatic feedback, whereas no statistically significant difference can be found between EXP and CTRL_NB for the phonemes for which no feedback was available, suggesting that the feedback provided by Dutch-CAPT was effective in improving the quality of the targeted phonemes and that training with automatic feedback at segmental level can offer added value with respect to

training without such feedback.

However, the fact that EXP made a significantly larger improvement than CTRL_NB on targeted phonemes may also be due to the presence of a floor effect for targeted errors around 11.5%, i.e. around the means of EXP and CTRL_NB at post-test. In other words, it is possible that the CTRL_NB participants could not achieve an improvement comparable to that of EXP on targeted phonemes because they did not start off at similar levels of errors and they could not possibly make fewer than 10.3% targeted errors. Similarly, the lack of a large improvement with respect to untargeted errors could be due to the fact that untargeted phonemes are less problematic and to the presence of another floor, this time for untargeted errors, around 3.5%. Given the initial difference between these two groups, and the fact that no data are available from comparable studies on possible floor effects, neither hypothesis can be ruled out on the basis of the analyses described above.

Nevertheless, if we look at the individual improvements of the subjects in EXP, we observe that three Arabic speakers who had lived in the Netherlands for many months were able to make large improvements (see Figure 6). Participant PA1, who had resided in the country for 10 months, showed a decrease in problematic errors of 7.25%. PA2, who had resided for 3.2 years in the Netherlands, showed a decrease of 24.7%. PA3, who had lived for 3.5 years in the Netherlands, achieved a decrease of 8%. Despite the typological distance of their mother tongue from Dutch and the very long time they had already been exposed to Dutch, these subjects show a strong improvement. This, together with the very different error behaviour they display for untargeted errors (see Figure 6), makes it reasonable to assume that the improvement of these subjects is
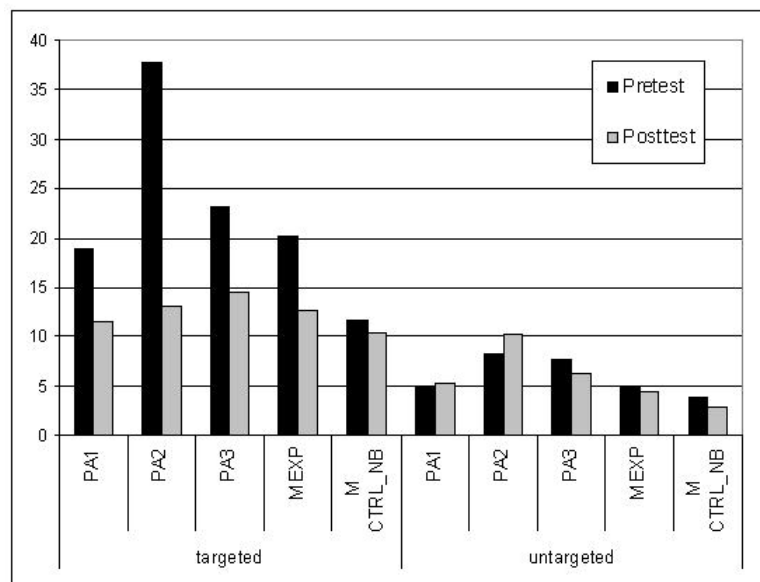

Fig. 6. Percentages of errors on targeted and untargeted phonemes at pre-test and post-test for individual subjects. The group means of EXP and CTRL_NB are also provided, for reference.

attributable to the feedback received. It would seem that training with Dutch-CAPT benefited these learners by accelerating their development of Dutch segmental quality.

5.3 Analysis of students' evaluations

The analyses of the answers given by EXP and CTRL_NB to two anonymous questionnaires about the specific CAPT system used reveal positive reactions to the two programs, which is congruent with results from several other studies and reviews of CAPT, including ASR-based CAPT (ALR, 1998; Harless *et al*., 1999; Mak *et al*., 2003). The answers indicate that the students enjoyed working with the CAPT system provided and that they generally believed in the usefulness of the training. Three of the fourteen participants who provided comments on *Dutch-CAPT* said that it was a good, interesting, and fun complement to the regular language course. Eight participants commented that it was helpful, mostly in making them aware of their specific pronunciation problems, but also in improving their vocabulary, their knowledge of spelling pronunciation issues in Dutch, and of the Dutch language in general. These comments were given in response to open-ended questions and seem to suggest that at least some learners in EXP were able to develop a metalinguistic awareness of the different phonemes of the Dutch language. For *Nieuwe Buren,* no such comments were made, even though the exercises in that system included tasks to train and test receptive discriminatory skills. Feedback on receptive skills may perhaps not be sufficient for learners to develop such an awareness.

The students who did not receive automatic feedback (CTRL_NB) seemed to find it a dispensable feature, whereas those who did receive it (EXP) seemed to believe quite strongly in its necessity. Possibly, the students in CTRL_NB did not notice certain errors in their pronunciation and thus simply assumed that theirs was a correct imitation of the target utterance, or they did manage to notice important errors simply by listening to their own recordings and by relying on their own discriminatory abilities. In contrast, participants in EXP might have become aware of the fact that they lacked such discriminatory skills, by studying the feedback they received, and they might have started to develop those skills. This seems corroborated by the fact that those participants generally found the feedback provided not detailed enough to improve their pronunciation.

One respondent who worked with *Nieuwe Buren* and four respondents who worked with *Dutch-CAPT* expressed a preference for minimal pair exercises, which simply required the participants to pronounce single words that they could hear and read on the screen. These exercises are considered useful to achieve certain articulatory dexterity and aural discrimination skills. At the same time, it is generally agreed that this kind of listen-and-repeat exercises should be integrated with less imitative speaking activities which are more realistic and effective to develop global communicative skills (Jones, 1997). The finding that some EXP participants liked these exercises the most might be explained by the fact that these exercises were 'easier': participants were able to focus on a single word and very few sounds each time, which generally meant that they succeeded in pronouncing the words and the sounds correctly, thus obtaining positive feedback. Exercises with longer sentences, especially when the sentences were not written out on the screen, were more challenging for these learners.

6 General discussion

The different analyses carried out in this study have provided us with information on the effectiveness of automatic corrective feedback on segmental quality and on CAPT in general. The ratings of global pronunciation quality indicated that the participants receiving automatic ASR-based feedback made larger improvements in segmental quality. However, the lack of a significant difference between the three groups' mean improvements suggests that the impact of corrective feedback on global segmental quality was limited. Possible explanations that were advanced include practical constraints in the experiment methodology, such as the small sample size and the impossibility of matching the groups prior to the training, as well as the small amount of computer-based training provided.

Global evaluations are a useful instrument to gauge the effectiveness of CAPT: they are ecologically relevant, because when learners are judged on their pronunciation skills in the real world, for instance during a job interview, they are judged on the basis of global impressions. However, if the resolution of the analysis is increased and specific errors are examined, as in our in-depth analyses, a clearer picture appears. Although it cannot be excluded that both groups of participants reached a floor with respect to the phonemes under investigation, the individual results obtained from these analyses seem to suggest that the CAPT system providing automatic feedback was effective in the task for which it was built, i.e. improving segmental quality on a selection of problematic phonemes. The ASR-based feedback seems particularly effective for learners who are generally lagging behind. These results are all the more encouraging since they were obtained with algorithms whose performance is not 100% error-free, as they suggest that the occasional erroneous feedback provided in *Dutch-CAPT* did not seriously hamper the learning process.

The more discrete evaluation of the phonemes addressed in the training is a fairer, more sensitive, and therefore more appropriate instrument to measure the effectiveness of this CAPT system and of the corrective feedback provided. This evaluation is particularly important for ASR-based CAPT systems, which are made of different components, so that possible problems can be ascribed to specific components and solved there. It is therefore advisable to adopt both types of analyses when evaluating these types of systems, so that both the specific effectiveness of automatic feedback and its ecological value can be established.

By combining the results of these two types of analyses, we can hypothesize that a few adjustments might suffice to obtain a stronger impact on global pronunciation quality. The training could be intensified. This could further reduce the number of segmental errors among the targeted phonemes and could provide the opportunity for generalizing knowledge acquired on certain phonemes and phonemic contrasts (e.g. on vowel length in Dutch) to other untargeted phonemes or contrasts, thus improving global segmental quality to a larger extent too. Moreover, the number of phonemes targeted by the feedback algorithm could be increased. Remember, however, that the feedback in this CAPT system was designed to be useful and robust for learners with different mother tongues: it is thus possible that the results obtained on global segmental quality in this study reflect the limitations of such an approach. A more targeted system specifically developed for speakers with the same mother tongue might be more

effective and show a stronger impact on global segmental quality. Such a solution would probably be at the expense of the usability of the system, however, which would be limited to a specific group of learners.

Linked to this point is another advantage of the combined analyses in this study: the discrepancy between scores for global segmental quality and scores for specific phonemes targeted in training can contribute to better understanding the relationship between local, discrete aspects of pronunciation quality and global ones (see Riney, Takada & Ota, 2000; Neri *et al*., 2006), which has yet to be clarified.

Finally, the questionnaire results indicate that the students in this experiment generally appreciated working with the CAPT systems they were assigned. They had a positive impression of this type of pronunciation training, finding it a useful and enjoyable complement to the regular lessons, a fact that should not be underestimated given the importance of motivational factors in language learning and, in particular, with respect to pronunciation skills (see Bongaerts, Van Summeren, Planken & Schils, 1997; Moyer, 1999). The students receiving automatic feedback on segmental quality found this feature necessary, tended to trust the system's feedback, and were generally satisfied with the implementation of this feature. Although this does not mean that the system could not be improved, the indication that participants trusted and liked the kind of corrective training provided in the system is an important, positive result because it means that learners will be willing to correct their pronunciation behaviour according to this feedback, and that they will be motivated to continue training in this way. At the same time, the answers to these questionnaires seem to suggest that automatic feedback on productive skills is more effective than feedback on receptive skills in developing a metalinguistic awareness of phonemic differences.

7 Conclusions

ASR-based CAPT has garnered considerable attention in the past decade for its possibility of providing corrective feedback on pronunciation, a task for which limited time is available in traditional instruction. The detailed analyses of segmental quality in this study suggest that the simple and easy-to-understand automatic feedback provided in our ASR-based CAPT system can be effective in improving segmental quality of the phonemes it targeted, leading to larger improvements than those achieved in the absence of such feedback, after only four hours of CAPT over a period of one month. In particular, this type of focused training seems to accelerate segmental development in learners who are generally slower. However, no statistically significant advantage appeared from a more global evaluation of segmental quality of the learners' speech. Several possible explanations for this finding have been proposed and discussed.

The positive results measured in all three analyses in this study, including the positive response of the students to the training and the feedback they received, give us reason to believe that a system providing simple, corrective automatic feedback can be a useful pedagogical tool to supplement regular teacher-fronted classes and that it is worthwhile to continue studying this technology so that it can be improved. The results of this study should be treated with caution, however, because of two main limitations. One has to do with the small sample size considered, and the heterogeneity of the groups due to the impossibility of matching subjects prior to the training. This implies that the results

obtained might not necessarily be generalizable. The other limitation has to do with the fact that this study did not investigate the long-term effects of the training, which means that no claim can be made with respect to the effectiveness of corrective feedback in the long term. Further, methodologically sound research into this type of application, with more balanced groups and with delayed post-tests, might confirm the trends evidenced in this study.

This study has nevertheless contributed to the research on ASR-based CAPT systems in at least two ways. First of all, by emphasizing the need for assessing these systems' effectiveness from the point of view of the actual improvements made by users. Secondly, by offering suggestions on how to go about investigating the effectiveness of these systems.

Acknowledgements

References

ALR (1998) Putting Pronunciation Programs Through Their Paces, *American Language Review*, 2. http://www.languagemagazine.com/internetedition/mj98/ eepp56.html

Bongaerts, T. (2001) Age-related differences in the acquisition of L2 pronunciation: The Critical Period Hypothesis revisited. Paper presented at the *EUROSLA preconference workshop on the Age Factor in L2 acquisition*, Paderborn, Germany.

Bongaerts, T., Van Summeren, C., Planken, B. and Schils, E. (1997) Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, **19**: 447-465.

Cucchiarini, 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.

De Bot, K. (1983) Visual feedback of intonation I: Effectiveness and induced practice behavior. *Language and Speech*, **26**: 331-350.

El-Tatawy, M. (2002) Corrective Feedback in Second Language Acquisition. *Working Papers in TESOL & Applied Linguistics*. Teacher's College, Columbia University. http://www.tc.columbia.edu/ academic/tesol/Webjournal/El-Tatawy.pdf

Felix, U. (2005) Analysing recent CALL effectiveness research. Towards a common agenda. *Computer Assisted Language Learning*, **18**: 1-32.

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

Hincks, R. (2005) Computer support for learners of spoken English. *Unpublished doctoral*

dissertation, KTH Stockholm, Sweden.

Jones, R. H. (1997) Beyond 'Listen and Repeat': Pronunciation teaching materials and theories of Second Language Acquisition. *System*, **25**: 103-112.

Kim, Y., Franco, H. and Neumeyer, L. (1997) Automatic pronunciation scoring of specific phone segments for language instruction, *Proceedings of Eurospeech*, Rhodes, Greece, 1997, 645-648.

Macdonald, D., Yule, G. and Powers, M. (1994) Attempts to improve English L2 pronunciation: the variable effects of different types of instruction. *Language Learning*, **44**: 75-100.

Mak, B. Siu, M., Ng, M., Tam, Y.-C., Chan, Y.-C., Chan, K.-W. *et al*. (2003) PLASER: Pronunciation Learning via Automatic Speech Recognition. *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications using Natural Language Processing*, Edmonton, Canada, 2003, 23-29.

Mayfield Tomokiyo, L., Wang, L. and Eskenazi, M. (2000) An empirical study of the effectiveness of speech-recognition-based pronunciation tutoring. *Proceedings of the 6th International Conference on Speech and Language Processing*, Beijing, China, 2000, 677-680.

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

Neri, A., Cucchiarini, C. and Strik, H. (2006) Selecting segmental errors in L2 Dutch for optimal pronunciation training. *International Review of Applied Linguistics*, **44**: 357-404.

Precoda, K., Halverson, C. A. and Franco, H. (2000) Effects of speech recognition-based pronunciation feedback on second-language pronunciation ability. *Proceedings of InSTIL 2000*, Dundee, Scotland, 2000, 102-105.

Purcell, E. T. and Suter, R. W. (1980) Predictors of pronunciation accuracy: A reexamination. *Language Learning*, **30**: 271-287.

Riney, T.J., Takada, M. and Ota, M. (2000) Segmentals and global foreign accent: The Japanese flap in EFL. *TESOL Quarterly*, **34**: 711-737.

Witt, S. M. and Young, S. J. (2000) Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, **30**: 95-108.