Pronunciation training in Dutch as a second language on the basis of automatic speech recognition

Ambra Neri, Catia Cucchiarini and Helmer Strik

Centre for Language and Speech Technology (CLST)

Radboud University Nijmegen, The Netherlands

Summary

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*, a Computer Assisted Pronunciation Training (CAPT) system with Automatic Speech Recognition (ASR) 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 (ASR) research 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 of pronunciation, which is considered one of the most difficult skills to learn in a second language. This is not surprising considering the advantages ASR-based Computer Assisted Pronunciation Training (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 (Ehsani & Knodt, 1998; Neri et al., 2002), it seems legitimate to question to what extent it is possible to develop ASR-based CAPT systems that provide feedback on pronunciation problems and that are indeed efficacious in improving pronunciation quality. To test this hypothesis we conducted a study in which we investigated which pronunciation problems can and should be addressed by an ASR-based CAPT system and whether training with such a system can lead to improvement of segmental quality.

2. Selection of pronunciation errors to be addressed

As many pronunciation errors are due to interference from the native language (L1), some CAPT systems target one specific language pair (e.g. Menzel et al. 2001; Mak et al. 2003) and are trained on a selection of typical L1-L2 errors. In this way, precise predictions can be made on possible errors, thus making the error identification procedure more efficient and boosting the effectiveness of the training. However, this approach is not always feasible, e.g. when no information is available on typical L1-L2 errors. Moreover, an L1-specific approach is not always desirable because the number of potential users will be limited to the speakers of the chosen language pair. In certain contexts, like the Dutch situation, it would not even be realistic because in Dutch-L2 classes learners with different mother tongues are grouped together.

To cater for such a heterogeneous group, within the framework of our research aimed at developing a CAPT system for Dutch L2, we examined available literature on pronunciation errors in nonnative Dutch finding that systematic studies

on this topic are scarce. Some information has been collected by speech therapists and language teachers within the Dutch-L2 teaching context (see: Aan de Wiel et al. 1991; Coenen, 1991; Tupker, 1995). Other studies considered the differences between Dutch and one or more L1s (e.g. De Graaf, 1986; Van Boeschoten, 1989; Doeleman, 1998).

For our purpose, however, we needed a detailed inventory of segmental errors in nonnative Dutch speech. For this reason, we analyzed three different databases, made annotations of errors they contained, carefully studied these annotations and selected a number of errors that, according to predetermined criteria, should first be addressed in a CAPT system.

2.1. Criteria for selecting pronunciation errors

To decide which errors should be addressed by the ASR-based CAPT system for Dutch we adopted the following six criteria: 1) common across speakers of various L1s; 2) perceptually salient; 3) potentially hampering to communication; 4) frequent; 5) persistent, and 6) automatically detectable with sufficient reliability.

First of all, considering the Dutch context, it was important to draw up an inventory that primarily contains errors shared by speakers with different mother tongues (criterion 1). Second, in line with the focus of current pronunciation training approaches, only realizations that are perceived by a human listener as clearly erroneous should be taken into consideration (criterion 2). Analogously, perceptually salient deviations that are unlikely to hamper communication, such as deviations that Dutch listeners are familiar with because they are found in certain well-known non-standard Dutch varieties, ought not to be prioritized (criterion 3). Furthermore, addressing infrequent errors is likely to have little impact on overall pronunciation performance, and on improving communication (criterion 4). In addition, it is wiser to spend time and effort in correcting errors that appear to persist over time and possibly over a certain amount of instruction, rather than errors that simply disappear through exposure to the L2 (criterion 5). Finally, since we know that the performance of ASR-based detection procedures is not 100% correct, it is important to avoid the risk of providing erroneous feedback and thus to choose only those pronunciation errors that appear to be detected automatically with sufficient reliability (criterion 6).

2.2. Three studies on the occurrence of pronunciation errors in non-native Dutch

The first database we studied consisted of 810 speech fragments read by 54 L2 learners of Dutch with 14 different L1s who were mainly at the intermediate and advanced proficiency levels. Read speech was chosen because it can ensure homogeneity in content, which in turn makes it more likely that annotators are not influenced by oral production factors beyond segmental quality. Moreover, using read speech makes it possible to select material with a complete phonemic makeup, which enables comparisons for all Dutch phonemes across speakers. Read speech has nevertheless also some disadvantages: it may be influenced by orthography to a larger degree than spontaneous speech (Young-Scholten, 1997) and the oral production obtained with read speech may not truly reflect the actual production of a learner engaged in a real-life speaking task.

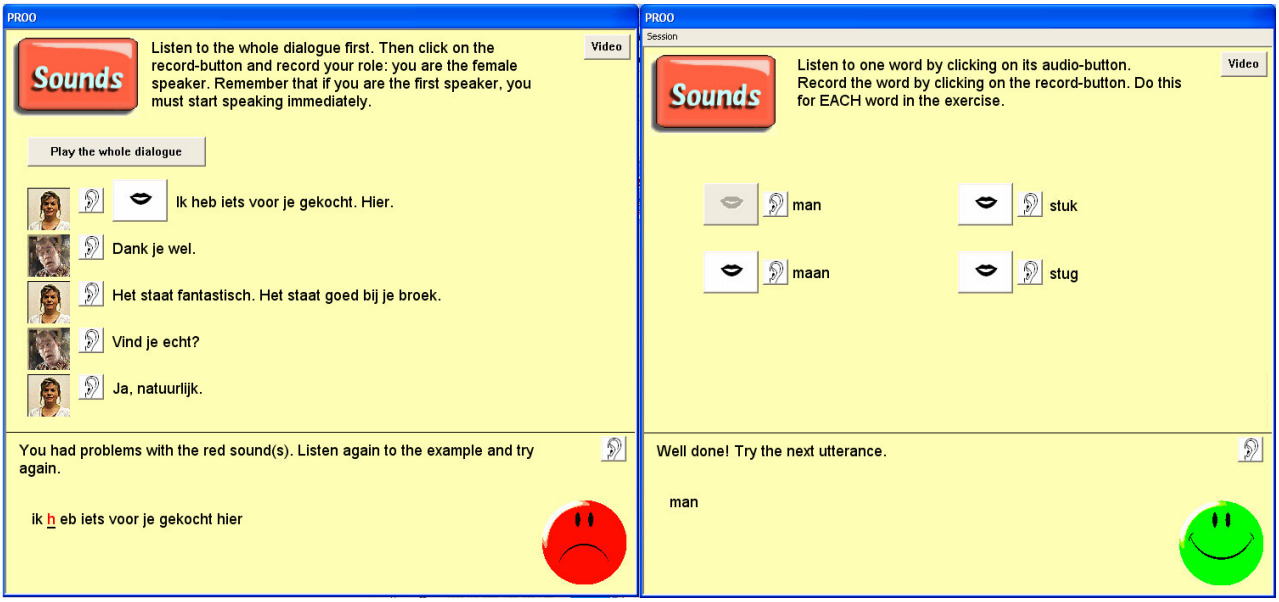
Consequently, to achieve greater robustness in our selection, we also studied 319 fragments of extemporaneous speech produced by 42 learners of Dutch who took part in a test of Dutch L2 proficiency. This database was chosen for its ecological validity, since it contains authentic samples of speech produced by real learners of Dutch who were taking a test.

Finally, we examined a database of read speech produced by Italian learners of Dutch as a foreign language (FL) who varied with respect to amount of instruction in Dutch. The speakers were 20 Italian students aged 19 to 25 attending a Dutch university course in Italy, who were selected across five years of instruction. This final analysis made it possible to compile a list of errors made in an FL context and to obtain more detailed information on persistent errors.

The procedure to obtain the final inventory of phonemic errors consisted in auditory analyses of the speech material, in which an expert annotated perceptually salient errors (for more details, see Neri et al., 2006) followed by quantitative and qualitative analyses aimed at refining the list and arriving at a robust final inventory of errors. In other words, only errors meeting all or most of the six criteria qualified to be included in the final inventory.

All three studies revealed that vowels are more problematic than consonants. Analyses of the three different databases of Dutch nonnative speech produced by a total of 116 learners with different mother tongues and proficiency levels based on the criteria outlined above led to the selection of the following eleven Dutch phonemes: /ɑ/, /aː/, /ɛi/, /ɪ/, /ʏ/, /y/, /u/, /œy/, /øː/, /h/, /χ/.

Fig. 1. Snapshot of a dialogue in which a phoneme was mispronounced in the first utterance (left) and of a minimal-pair exercise (right).



3. ASR-based CAPT system

The ASR-based CAPT system used in this study, which we call *Dutch-CAPT*, is a computer program developed at our department that provides feedback on Dutch pronunciation. For the content, we built on *Nieuwe Buren* (New Neighbours), a comprehensive CALL program used by schools for Dutch L2 in the Netherlands and designed specifically for literate adult L2 learners with different L1s. The exercises in *Dutch-CAPT* include role-plays, questions to be answered by recording one of several possible answers, and exercises requiring students to pronounce words and minimal pairs for which example pronunciations are given. The program provides feedback on the eleven Dutch phonemes indicated above.

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 utterance pronounced by the student together with a smiley and a short written comment. If the ASR algorithm finds that a phoneme has been mispronounced, the letter(s) corresponding to the mispronounced phoneme is underlined and 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 (Fig. 1, left), and the student is prompted to repeat the utterance. In this way the feedback is simple and concise, and clearly corrective. 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. If the learner's utterance is correct, a green smiley appears instead (Fig. 1, right).

4. The effectiveness of ASR-based feedback

To establish the effectiveness of *Dutch-CAPT* in realistic conditions, we studied a group of immigrants who were learning Dutch in the Netherlands. 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.

4.1. Subjects

The participants were 30 adult immigrants varying with respect to age, occupation and length of residence in the Netherlands who were following beginner courses of Dutch at UTN, the language centre of the Radboud University Nijmegen. They were 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:

- Experimental group (EXP). Fifteen participants, 10 female and 5 male.
- Control group 1 (NiBu). Ten (4 female and 6 male) participants.
- Control group 2 (noXT). Five (3 female, 2 male) participants.

4.2. Training procedure

All three groups followed the regular classes, which included self-study sessions in the language lab. NiBu and EXP had one 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* 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 of each student's activities allowed the experimenter to check that all students completed all exercises as requested. EXP used *Dutch-CAPT*, which was similar to *Nieuwe Buren*, the only difference being that it provided automatic feedback on segmental quality.

4.3. Testing procedure

4.3.1. 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 1-5 Likert scale and to answer 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.

4.3.2. 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).

Two groups of three expert raters evaluated the speech independently on a 10-point scale, where 1 indicated very poor segmental quality 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. 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 (Cucchiarini et al. 2000). Pre- and post-test recordings were presented in random order.

4.3.3. 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 on a subset of the participants' recordings, and annotations were made of specific segmental errors.

4.4. Results

4.4.1. Analysis of students' evaluations

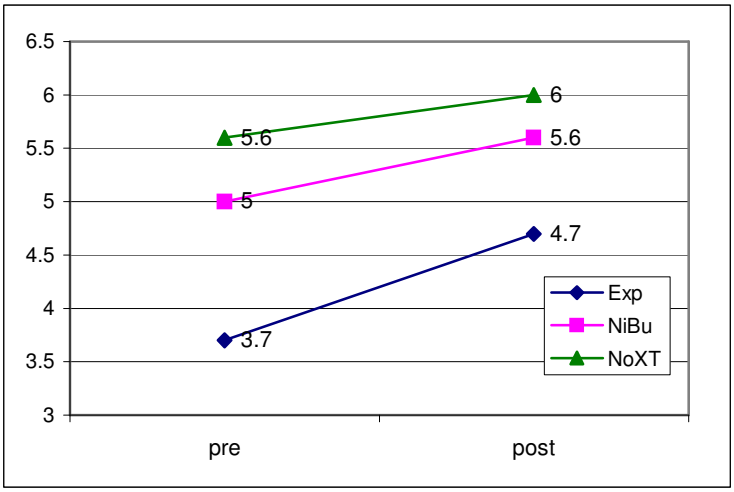
Overall, the responses indicated a positive reaction to the two CAPT programs, with mean scores per statement ranging between 2.4 and of 4.6 for EXP, and between 2.3 and 4.7 for NiBu. This result is in line with other studies on student appreciation of CAPT, including ASR-based CAPT (Mak et al., 2003). 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.

4.4.2. 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. For the two groups of raters 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, pre-specified criteria for assessment were provided.

After having established that the non-natives had not received scores in the range of the natives at pre-test, we 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. 2). Subsequently, an analysis of variance (ANOVA) with repeated measures was conducted. 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.

Fig. 2. Mean scores (based on a 10-point scale) for each training group at each testing condition.



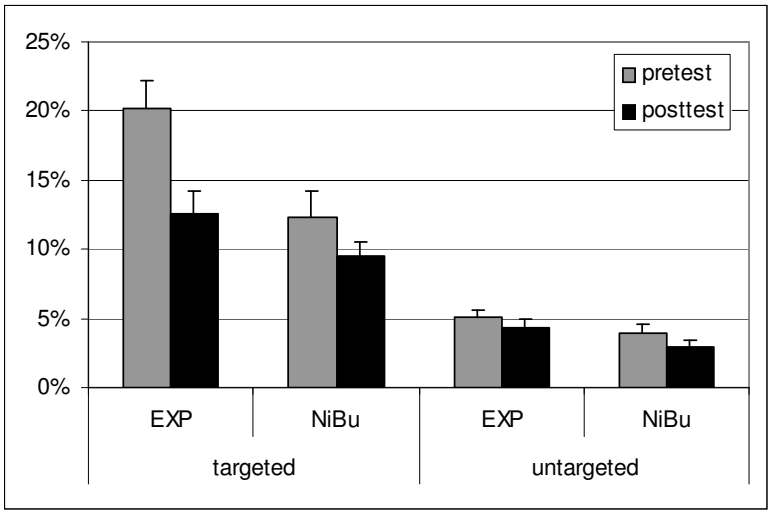
To summarize, all three groups improved global segmental quality after the training, with the group receiving ASR-based corrective feedback showing the largest improvements, followed by the group receiving extra CAPT, but no corrective feedback. However, the difference in improvements in the three groups 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, which resulted in EXP's and noXT's mean scores being significantly different at pre-test. To better understand these results, we carried out a fine-grained analysis of the segmental errors made by the participants before and after the training.

4.4.3. In-depth analysis of segmental quality

An expert annotator listened to the recordings and indicated whenever a phoneme was pronounced incorrectly. We used these annotations to examine possible post-test improvements on the 11 target phonemes and on the remaining, untargeted phonemes for the participants who received automatic feedback and those who did not. To obtain two comparable groups, we removed noXT from these analyses and retained EXP and NiBu, which differed from each other only in the variable 'feedback'.

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. The results (see Fig. 3) show first of all that there were more errors on the targeted phonemes than on the untargeted ones. Second, it appears that errors on the targeted phonemes decreased by 7.6% (absolute decrease, SD = .074) for EXP and by 1.4% (SD = .029) for NiBu. As these data did not meet all assumptions necessary for an ANOVA, non-parametric tests were performed. A Wilcoxon Signed Ranks test on the subjects' scores for targeted and untargeted errors indicated an overall improvement on both types of errors at post-test for all subjects: 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. The test on targeted errors indicated a significant difference between EXP and NiBu (z=-2.827, p = .002, one-tailed), with EXP making a significantly larger improvement than NiBu on segmental quality of the targeted phonemes.

Fig. 3. Mean error percentages (and *SEM*s) for errors on the targeted and untargeted phonemes.



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 (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 Fig. 3): While both groups produced fewer errors at post-test, the decreases in untargeted errors were much smaller and more similar across the two groups (0.7% for EXP and 1.1% for NiBu) than those for the targeted errors. The Wilcoxon Rank-sum test on untargeted errors indicates no significant difference between the mean improvements of the two groups on untargeted phonemes.

In summary, these results show that a) the participants produced relatively 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 than the group receiving no automatic feedback, 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 and that training with automatic feedback at segmental level can offer added value with respect to training without such feedback.

5. Discussion and conclusions

The study on the feasibility and the effectiveness of ASR-based *Dutch-CAPT* presented in this paper has shown that a) it is possible to develop a system that provides automatic feedback on a number of Dutch sounds that appear to be problematic for learners with different mother tongues, b) that the students enjoyed using this system and c) that this seems to be 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 emerged 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 better understand 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. Various factors might be responsible for these findings.

First of all, the fact that the training concerned a limited number of phonemes that were problematic for a heterogeneous group of learners. This means that for each specific learner very few problematic phonemes could be addressed. It is therefore possible that a different approach in which individual learners could receive feedback on a larger number of problematic phonemes would have been more effective at the global level. Such an approach could either consist in having a system that simply addresses more phonemes, or a system that is specifically designed for addressing pronunciation errors pertaining to specific language combinations such as, for instance, Turkish-Dutch or French-Dutch. Such an L1-specific CAPT-system might be more effective. However, the question then is how feasible it is to develop a variety of such systems for all L1s that might be relevant in a given linguistic context.

A second factor that might explain why the impact of the training was not significant at the global level is that the training was very short and non-intensive. This might have prevented the learning effect from generalizing to other, similar phonetic contrasts, for instance that between /oː/and /ɔ/and that between /eː/ and /ɛ/ as a result of training the /aː/-/ɑ/ contrast. These are issues that we intend to address in future research.

6. Acknowledgements

The present research was supported by the Dutch Organization for Scientific Research (NWO). We would like to thank Ming-Yi Tsai, F. de Wet, M. Hulsbosch, L. ten Bosch, C.van Bael, J. Kerkhoff, and A. Russel for their help building

Dutch-CAPT, and L. Boves and T. Rietveld for their valuable comments on the analyses presented in this paper. Many thanks also go to the students and teachers at UTN. Finally, we are indebted to Malmberg Publishers for giving us the opportunity of using the educational program *Nieuwe Buren* for our research.

7. References

Aan de Wiel, M.C., Van den Brink, G.A., Struijk van Bergen, S.G.E. (1991). Diagnostiek van uitspraakproblemen van tweede-taalverwervers van het Nederlands. *Levende Talen* 466, 507-511.

Van Boeschoten, J.A. (1989). *Verstaanbaarheid van klanken in het Nederlands gesproken door Turken*. Dordrecht: ICG Printing.

Coenen, J. (1991). *Uitgesproken Nederlands*. Groningen: Wolters-Noordhoff.

Cucchiarini, C., Strik, H., & 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.

Doeleman, R.(1998). *Native reactions to nonnative speech*. Doctoral dissertation. Tilburg: Tilburg University Press.

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

Graaf, de T. (1986). De uitspraak van het Nederlands door buitenlanders. *Logopedie en Foniatrie,* 58, 343-347.

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

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

Menzel, W., Herron, D., Morton, R., Pezzotta, D., Bonaventura, P., & Howarth, P. (2001). Interactive Pronunciation training. *ReCALL,* 13, 67-78.

Tupker, Y. (1995). *Uitspraak Onderwijzen*. Manuscript. Tilburg: Katholieke Universiteit Brabant.

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

Neri, A., C. Cucchiarini, and H. Strik (2006). Selecting segmental errors in non-native Dutch for optimal pronunciation training. *International Review of Applied Linguistics in Language Teaching*, 44, 357-401.

Young-Scholten, M. (1997). Second-language syllable simplification: Deviant development or deviant input? In New Sounds 97: *Proceedings of the Third Symposium on the acquisition of Second-Language Speech*, J. Leather and A. James (Eds.), 351-360. Klagenfurt: University of Klagenfurt.