ASR-based pronunciation training: Scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners

Catia Cucchiarini [1], *Ambra Neri* [1], *Febe de Wet* [2], *Helmer Strik* [1]

[1] CLST, Department of Linguistics, Radboud University, Nijmegen, The Netherlands
[2] SU-CLaST, Stellenbosch University, South-Africa

[c.cucchiarini|a.neri|h.strik]@let.ru.nl, fdw@sun.ac.za

Abstract

A system for providing Computer Assisted Pronunciation Training for Dutch was developed, Dutch-CAPT, which appeared to be effective in improving pronunciation quality of L2 learners of Dutch. In this paper we describe the architecture of the system paying particular attention to the rationale behind this system, to the performance of the error detection algorithm and its relationship to the pedagogical effectiveness of the corrective feedback provided

Index Terms: Computer Assisted Pronunciation Training (CAPT), corrective feedback, pronunciation error detection, Goodness Of Pronunciation (GOP)

1. Introduction

In [1] we reported on a study aimed at determining the pedagogical effectiveness of a specific application of Computer Assisted Pronunciation Training (CAPT) that we had developed for Dutch, Dutch-CAPT. This system provides corrective feedback on a selected number of speech sounds that have appeared to be problematic for learners of Dutch from a variety of L1 backgrounds [2; 3]. The results in [1] showed that for the experimental group that had been using the CAPT system for four weeks, the reduction in the pronunciation errors addressed in the training system was significantly larger than in the control group. Although these results are very promising and open up new avenues for pronunciation training in Dutch L2, it is to be expected that the algorithm for pronunciation error detection employed in this system was not foolproof, but now and then made mistakes, as is usually the case with such algorithms [4]. It is therefore interesting to find out how accurate the algorithm was in detecting mistakes because this can provide insight into the relationship between accuracy in error detection and effectiveness of a system and it can increase our understanding of the possibilities of CAPT and the potential to employ this technology in L2 learning.

In this paper we first describe the architecture of the system and provide details on the decisions underlying the specific approach adopted in building this system (section 2). Subsequently, in section 3 we pay attention to the implementation of the error detection algorithm, in section 4 to its performance and in section 5 to the effectiveness of Dutch-CAPT. In section 6 we discuss the results and in section 7 we present some conclusions.

2. Functional description of Dutch-CAPT

Dutch-CAPT is a computer program developed at the Radboud University in Nijmegen 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 Dutch-CAPT system is comprised of two main parts, a client and a server. The user interface (UI), which includes the didactic content of the system, is on the client side of the system. The server contains the technology that analyses the students' utterances, including the ASR module. The two parts communicate through two sockets, one to exchange commands, the other to exchange speech data (Figure 1).

2.1. The client

The client contains the UI and the didactic content of Dutch-CAPT. An English and a Dutch version of the UI are available. When the client is started, the user is prompted to enter a personal four-digit ID, to indicate his/her gender and whether s/he prefers to use the English or the Dutch version of the instructions. The ID is needed to keep logs of the students' activities and the gender has to be specified because the server makes use of different parameter settings for the acoustic analyses of male and female speakers.

The navigation through the exercises is constrained and sequential and requires users to complete an exercise before proceeding to the following one. This ensured that the subjects received the same amount of training, as the system was built for research purposes. The constrained navigation can nevertheless be overruled by the experimenter.

The didactic content consists of 106 exercises grouped into four units each containing a video providing typical communicative situations (buying groceries, going to the cinema etc.) with words and expressions to be practised orally in that unit and different exercises with which users can practise pronunciation. They include:

- 22 exercises in the form of dialogues. These exercises simulate parts of the dialogues presented in the videos. The user has to choose one role and to pronounce the character's lines in a flow, as if actually speaking with the other character.
- 46 exercises consisting of questions that are either written out or are only available in audio-format.
- 38 exercises consisting of individual words that have to be pronounced on the basis of the model utterances recorded by male and female native speakers. These exercises include several minimal pairs.

2.2. The server

The server contains the technology that performs the analyses on the users' utterances and is able to handle multiple simultaneous client processes, for which it creates separate IDs. A log file is maintained in the server, which contains a list of important tasks happening both in the client(s) and in the server. These tasks are provided with

unique process IDs and time stamps, e.g. the start of the recognition of an utterance or the results of the analysis on its quality. This log file is saved on the server, while a copy is simultaneously updated on a website, so that the experimenter can monitor several users in real time.

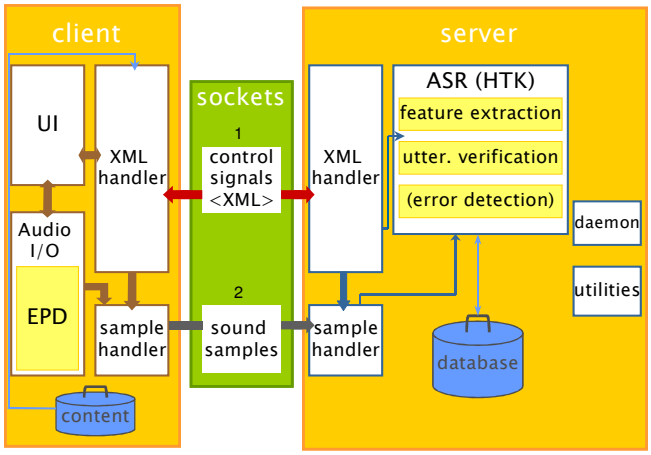

Figure 1 Schematic representation of Dutch-CAPT

2.3. How Dutch-CAPT works

The exercises are presented on the top half of the screen and the feedback on the lower half. For each utterance example pronunciations are available that can be listened to before recording an answer. When a user starts recording the answer, a so-called Endpoint Detector is started so that the recording can be automatically stopped once the user has finished speaking. The server must first of all establish whether the audio file received matches one of the possible answers for that given exercise, for two reasons. One has to do with the credibility of the system: If a user can utter any word or sentence including, for instance, 'I am tired of using this system' without the system noticing that this is not one of the correct options, s/he will soon stop trusting the system as a useful tool. Though this observation may sound obvious, this problem is a serious one in CAPT systems providing automatic feedback without ASR technology. The other reason is a practical one: If it cannot be established which utterance the user has pronounced, it is also impossible to provide information on the words and sounds in that utterance. In such cases, the system will reject the utterance and prompt the user to try again. If the server finds a suitable match to the user's audio file, it immediately starts analyzing pronunciation quality. If no error is found, (a) a message congratulating the student, (b) the orthographic transcription of the utterance recognized, and (c) a green, happy smiley will appear on the lower half of the UI together with a play button enabling the user to listen again to his/own pronunciation.

If an error is found, the client will display (a) a message informing the student that there was a problem with one or more sounds and prompting him/her to try again after listening to the target pronunciation, (b) the transcription of the utterance with the grapheme(s) corresponding to the erroneous phoneme(s) coloured red and underlined, and (c) a red, disappointed smiley (see Figure X). No more than three errors are signalled each time in order not to discourage the student.

If the user does not succeed in pronouncing a target utterance correctly over three successive attempts, a message indicating that there are still some problems is shown and the student is allowed to move to the following exercise.

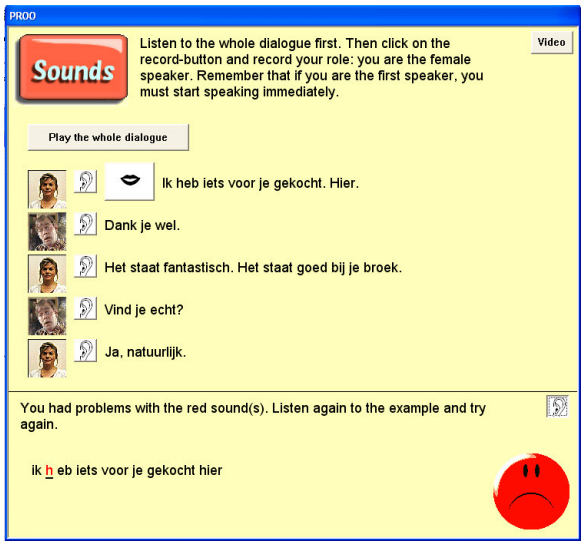

Figure 2 Screenshot taken after the female user has received negative feedback

3. Error detection in Dutch-CAPT

In accordance with the practice in Dutch L2 courses, which contain heterogeneous groups with respect to mother tongue, Dutch-CAPT had to address pronunciation errors that can be made by any learner, regardless of his/her L1. In addition, the rationale behind this system was that it did not have to address all pronunciation errors, but only a selection of relevant ones. To decide which errors should be addressed by Dutch-CAPT we adopted the following five criteria:

1. Common across speakers of various L1s
2. Perceptually salient
3. Potentially hampering to communication.
4. Frequent
5. Persistent

The first criterion is obvious given the Dutch context. The second criterion is in line with current pronunciation training approaches that tend to focus on realizations that are perceived as clearly erroneous by human listeners. Along the same lines, deviant realizations that Dutch listeners are familiar with because they are found in certain well-known non-standard Dutch varieties ought not to be prioritized, whereas attention should be concentrated on perceptually salient deviations that are likely to hamper communication. Furthermore, to improve communication it is important to address frequent errors (criterion 4) and errors that appear to persist over time (criterion 5). To obtain a detailed inventory of segmental errors in non-native Dutch speech, we analyzed three different databases of Dutch non-native speech produced by a total of 116 learners with different mother tongues and proficiency levels, made annotations of perceptually salient errors, carefully studied these annotations and selected a number of errors that, according to the predetermined criteria, should be addressed in Dutch-CAPT. These analyses led to the selection of the following 11 Dutch phonemes: /x/ /y/ /h/ /øː/ /ɑ/ /aː/ /ɛi / /œy/ /ɪ/ /ʏ/ /u/.

The ASR module was implemented in HTK and made use of 37 context-independent, monophone Hidden Markov Models (HMM). These HMMs were gender-dependent and were trained on read, native speech from the library of the blind and the broadcast news of the Spoken Dutch Corpus [5]. The phone set included a general speech model to account for unintelligible speech as well as a silence and a short pause model. Except for the short pause model, each HMM had three states and 32 Gaussian mixtures per state. The single state of the short pause model was tied to the central state of the silence model. Optimal Word Insertion penalty (WIP) and Language Model factor (LMF) values were determined using a development test set. The acoustic pre-processing implemented in the current version of Dutch-CAPT includes channel normalization, to account for the effect of using different microphones.

The error-detection algorithm was implemented by means of confidence measures (CMs) according to the method proposed in [6; 7]. Each utterance was subjected to a free phone and forced recognition using the HMMs described above. A so-called Goodness of Pronunciation (GOP) score was subsequently derived at phone level for the 11 Dutch phonemes selected. The GOP score for each phoneme corresponds to the ratio between the log-likelihood scores of its forced recognition and free phone recognition, normalized at the frame level. The lower the GOP score, the better the phoneme quality: If the GOP score of a specific phone falls below a certain threshold, it is accepted as a correct instance of the phone and vice versa. As in [6] thresholds per phone were obtained by using native speech material in which errors had been artificially introduced. In our approach we introduced errors according to patterns of substitutions deletions and insertions that we had observed in non-native speech [2; 3].

Target phones could be classified as:

- Correct Accept (CA) - A correct instance of a target phone is classified as correct by the CMs;
- False Reject (FR) - A correct instance of a target phone is classified as incorrect by the CMs;
- Correct Reject (CR) - An incorrect instance of a target phone is classified as incorrect by the CMs;
- False Accept (FA) - An incorrect instance of a target phone is classified as correct by the CMs.

The optimisation criterion used here consisted in maximising scoring accuracy, established by the following formula:

$$\text{Scoring Accuracy} = 100 \times \left(\frac{\text{Correct Acceptances} + \text{Correct Rejections}}{\text{Total number of target phonemes}} \right)$$

for a FR rate below 10%. This criterion was motivated by the fact that the users of the system were beginners. For these users, avoiding erroneously rejecting correct sounds was considered more important than erroneously accepting incorrect ones.

4. Scoring accuracy in Dutch-CAPT

The scoring accuracy of the algorithm was measured for a total of 2174 phones deriving from 437 utterances produced by the 15 Dutch immigrants who used Dutch-CAPT. These utterances were selected semi-randomly: For each participant, a maximum of 30 utterances was selected which contained at least one error as detected by the GOP algorithm. A Dutch expert annotator carried out auditory analyses based on the system's output. She was asked to indicate if she disagreed with the algorithm by indicating whether the errors identified were actually correct realizations (FR) and whether phonemes that had been seriously mispronounced had not been identified by the algorithm (FA). The tagged orthographic transcription was used as a starting point in order to make the task more efficient. The results indicate a mean scoring accuracy of 86% (SD = 4%). The results per subjects are in Table 1. As can be seen from this table, the percentage of FR stays below 10%.

Table 1. Algorithm's classification accuracy per subject.

Subject	N phones	N target phones	SA (%)	FR (%)
1	424	130	83	7
2	386	108	83	8
3	480	148	85	3
4	500	146	89	3
5	462	122	84	4
6	680	170	86	5
7	413	125	88	6
8	448	120	90	3
9	398	133	74	6
10	512	156	79	7
11	412	130	88	2
12	481	141	90	4
13	485	137	87	9
14	591	164	91	6
15	847	244	89	1
Total	7519	2174	86	5

5. Effectiveness of Dutch-CAPT

For the sake of completeness, we show here the results concerning the effectiveness of the feedback provided by Dutch-CAPT in improving problematic pronunciation errors that are reported on in [1; 10].

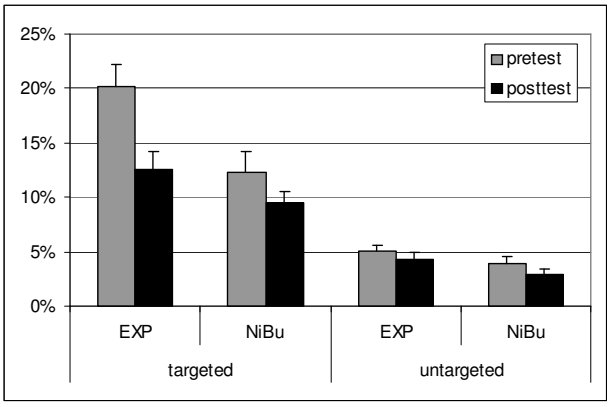

Figure. 3. Mean error percentages (and *SEM*s) for errors on the targeted and untargeted phonemes.

This figure shows that the participants produced relatively more errors on the targeted phonemes, which confirms that

these phonemes are, indeed, particularly problematic. In addition, it is clear that the group receiving feedback on these errors (EXP) made a significantly larger improvement on the targeted phonemes than the control group who used the same system but received no automatic feedback (NiBu), whereas no statistically significant difference was found for the phonemes for which no feedback was provided. This suggests that the automatic feedback provided in Dutch-CAPT was effective in improving the quality of the targeted phonemes.

6. Discussion

The data on the performance of the algorithm used for error detection in Dutch-CAPT that are presented in section 4 reveal that the scoring accuracy was relatively good and comparable to results obtained previously [7]. However, it is clear that scoring accuracy was not hundred percent. Nevertheless, training with Dutch-CAPT managed to significantly improve the pronunciation of notoriously problematic sounds, as shown in section 5, even for speakers of languages that are typologically distant from Dutch such as Arabic and Turkish, who are known to have problems learning Dutch pronunciation [10]. Similar results were obtained by [11].

There are a number of considerations underlying Dutch-CAPT that are worth examining when trying to understand why Dutch-CAPT was successful even though the performance of the algorithm was not perfect. First of all, the decision to provide feedback only on a limited number of well-selected problematic sounds. Second, the decision to limit the feedback to a maximum of three errors per utterance so as to avoid overwhelming the learners with too much information on many different mistakes. Third, the simple and straightforward feedback provided. Research on the effectiveness of feedback in L2 teaching has revealed that providing examples of the mistakes produced is probably unnecessary, while it suffices to point out the problematic areas. The type of corrective feedback provided in Dutch-CAPT comes very close to what is considered to be optimal corrective feedback "clear enough to be perceived as such" and allowing for self-repair and modified output [12]. A final point that probably deserves attention is the decision made in Dutch-CAPT with respect to the balance between FAs and FRs. As is well known, there is a trade-off between FAs and FRs. In Dutch-CAPT we decided to minimize FRs and tolerate FAs on the grounds that for beginner learners erroneously rejecting correct sounds would be more detrimental than erroneously accepting incorrect ones. As a matter of fact, beginners are likely to produce many errors and, accordingly, to receive a considerable amount of negative feedback. To avoid the risk of frustrating the students by rejecting correct utterances too [8; 9] it seemed wiser to tune the error detection algorithm so that the chance of false rejects would be as low as possible. In other words, in Ducth-CAPT only patently wrong sounds were (correctly) rejected, while a number of incorrect sounds were accepted as correct. This might have had a positive effect by enabling the learners to concentrate only on the most serious errors and to gain self-confidence [8].

7. Conclusions

In this paper we have described Dutch-CAPT, a system for providing automatic corrective feedback on pronunciation errors in Dutch, focusing particularly on error detection, scoring accuracy and feedback effectiveness. We have seen that a system that does not achieve 100% accuracy in error detection can still be effective in improving pronunciation errors, provided that the right decisions are made with respect to the implementation of corrective feedback.

8. Acknowledgements

The present research was supported by the Dutch Organization for Scientific Research (NWO). We are indebted to Ming-Yi Tsai, M. Hulsbosch, L. ten Bosch, C.van Bael, J. Kerkhoff, and A. Russel for their help building Dutch-CAPT , to L. Aul for the analyses of scoring accuracy and, finally, to Malmberg Publishers for giving us the opportunity of using the educational program Nieuwe Buren for our research.

9. References

[1] Neri, A., Cucchiarini, C. and Strik, H., "ASR corrective feedback on pronunciation: Does it really work?", Proceedings of Interspeech, 1982-1985, 2006a.

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

[3] Neri, A, Cucchiarini, C., Strik, H.. Selecting segmental errors in non-native Dutch for optimal pronunciation training. *International Review of Applied Linguistics*, 44, 2006b.

[4] Kim, Y., Franco, H. and Neumeyer, L., "Automatic pronunciation scoring of specific phone segments for language instruction", Proceedings of Eurospeech, 645-648, 1997.

[5] Oostdijk, N. The design of the spoken Dutch corpus. In Peters, P., Collins, P., and Smith, A., (eds.) New Frontiers of Corpus Research, Rodopi, Amsterdam, 105–112, 2002.

[6] Witt, S.M., Use of speech recognition in Computer-assisted Language Learning, Phd thesis, Department of Engineering, University of Cambridge, 1999.

[7] Witt, S.M. and Young, S., "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning", Speech Communication 30, 95-108, 2000.

[8] Egan, K. Speaking: A Critical Skill and Challenge. *CALICO Journal*, 16. 277-293, 1999.

[9] Herron, D., Menzel, W., Atwell, E., Bisiani, R., Daneluzzi, F., Morton, R. Automatic localization and diagnosis of pronunciation errors for second-language learners of English, Proceedings of Eurospeech '99, Budapest, Hungary, 855-858, 1999.

[10] Neri, A. The pedagogical effectiveness of ASR-based Computer Assisted Pronunciation Training, Phd thesis, Radboud University Nijmegen, 2007.

[11] Mayfield Tomokiyo, L., Wang, L., and Eskenazi, M., An Empirical Study of the Effectiveness of Speech-Recognition-Based Pronunciation Training, Proceedings ICSLP 2000, Beijing, China

[12] El Tatawy, M. Corrective feedback in second language acquisition, *Working papers in TESOL and Applied Linguistics*, Vol. 2. No. 2, 1-19, 2002.