Published in: Jokisch, Oliver (ed.): Elektronische Sprachsignalverarbeitung. Tagungsband der 27. Konferenz Leipzig, 2. – 4. März 2016. - Dresden: TUDpress, 2016. Pp. 175-182. (Studientexte zur Sprachkommunikation 81)

ON THE ASSESSMENT OF COMPUTER-ASSISTED PRONUNCIATION TRAINING TOOLS

Jürgen Trouvain[1], Jeanin Jügler[1] and Yves Laprie[2]

[1]Computational Linguistics and Phonetics, Saarland University, Saarbrücken (Germany)
[2]LORIA, Nancy (France)
trouvain@coli.uni-saarland.de

Abstract: The aim of this study is to select and formulate criteria for the assessment of tools and exercises that are using computer-assisted pronunciation training (CAPT). We examined ten different CAPT tools selected on the basis of an informal questionnaire among 10 colleagues working in a German-French CAPT project. Although the applied assessment must still be regarded as informal, and although the selected CAPT tools might not be an optimal sample for representing the state of the art, the results clearly show that there is a lot to improve regarding the clarity of instruction, the quality of exercises, the robustness of the diagnosis, the clarity and appropriateness of scoring, the diversity of feedback methods, the assumed benefit for various types of users as well as the usage of ASR. Despite various good approaches regarding graphics and game-like exercises there are obviously missing links between the pedagogical expertise in phonetic training on the one hand, and software development including usability engineering on the other.

1 Introduction

Computer-assisted language learning (CALL) and particularly computer-assisted (or computer-aided) pronunciation training (CAPT) can be considered as an optimal complement to a teacher-based class but also as the core technology of an entire language learning course. It can help to improve speech perception and speech production skills by raising awareness for phonological contrasts through exercises with discrimination tasks and those involving the user's own speech with the help of automatic speech recognition (ASR). Although there is a high potential for speech technology to be used in language learning scenarios there are also critical views on CAPT.

The aim of this study is to select and formulate criteria for the assessment of tools and exercises that are using CAPT. The selection of these criteria was developed through an evaluation of existing CAPT tools.

Although a CAPT tool is not meant to replace a teacher it carries out functions similar to those of a competent teacher and recommended teaching material – to some limited extent. Pronunciation training in classical classroom instruction consists of two bases. First, the interaction with a teaching person, which most of the time means to *listen* to a model speaker, interspersed with own speaking activities from time to time. Second, and this can be seen as an addition to teacher-interaction, exercises that are performed either at home or in school by the individual learner. These exercises are usually based on written material (books, exercise sheets) and sometimes recorded audio material is included. Sometimes also dialogue situations with interactional partners are practised.

In an "ideal world", CAPT would adopt the best practises of class-room pronunciation training which is teacher-assisted, book-assisted, audio material-assisted and dialogue partner-assisted. Although we have a limited reliable knowledge about which methods with which material are best for pronunciation training (PT), there are some ideas about the competences of an "ideal teacher" for PT [3]. S/he can provide explanations and can select exercises that

are appropriate for the learner considering factors like age, first language (L1), learning and cultural background and motivation. Learners with a non-alphabetical background of writing who aim for better reading and oral skills need a different training than people who want to learn some basics of a foreign language (L2) for touristic reasons, which is different again to university students learning an L2 in their home countries.

Another important competence of CAPT is the evaluation of errors and rendering assistance in their improvement. Errors of L2 speakers can be heavy or mild, they can concern an important phenomenon or can be more cosmetic, and the source can be purely orthography-based or can be traced back to the phonology of the L1 of the learner. Learners are often not aware of these distinctions. At any rate, the ideal teacher would carefully monitor the individual progress of the learner and would support the learner at the different stages of improvement, for learners some problems in PT are harder to overcome than others. Giving feedback is one of the central activities and functions of the teacher in PT. In the ideal case, the ideal teacher gives individual PT.

CAPT is always individual PT. Although forms of PT that are outlined above cannot be copied in a simple way to CAPT, the principles of PT in general can be kept in mind when developing and when assessing CAPT tools.

2 Method

Former studies of CAPT assessment were either based on pre- and post-tests of experimental and control groups of learners [4,5] or on self-assessment of an own system [2]. For this study we used an informal survey of experts who tested more than one system. We examined ten different (commercial and semi-commercial) CAPT tools selected on the basis of an informal questionnaire among 10 colleagues working in a German-French CAPT project. The software ranges from dedicated pronunciation trainers to CAPT exercises in CALL courses.

The tools were (in alphabetical order): AzAR, Babbel, Busuu, Duolingo, Cool Speech, English Central, Eye Speak English, Pronunciation Coach, Rosetta Stone, Speak Greek. Target languages differed for the testing, among them were English, German, Greek and French. All but one tool could be tested on a PC, the exception was tested on a tablet. The mean time the informal testers spend with one tool was between thirty and sixty minutes.

The assessment was performed assuming a beginning L2 learner as user of the CAPT tool in mind. In some cases native speakers tested software to assess the feedback robustness and to evaluate whether good pronunciations realised by native speakers were correctly identified. The following questions divided in various areas were meant to guideline the assessment:

- Level of proficiency and other individual needs
    - Does the tool consider the level of proficiency (and further characteristics of the learner) for the training programme?
    - Does the tool consider the level of further characteristics of the learner (age, motivation, L1) for the training programme?
    - Is there a placement test that considers pronunciation proficiency?
    - Are the exercises tailored to a specific L1?
- Instructions
    - Are the instructions clearly formulated?
    - Is it clear why a specific exercise should be performed?
    - How easy is the orientation in the training programme?

    - How feasible is the task to be performed?
- Quality of the exercises
    - Are there enough exercises for PT?
    - Do the exercises cover the most important topics of pronunciation problems in L2 learning?
    - Is prosody and fluency explicitly considered in PT?
- Types of exercises
    - Do the exercises vary in their type?
    - Are there any exercises that are not based on ASR or other methods of automated speech analysis?
    - Are there listening exercises with a focus on pronunciation, prosody and fluency, not only on understanding?
    - Is orthography in terms of letter-to-sound rules considered in the exercises?
- Feedback
    - Is there any corrective feedback beyond a simple "correct-incorrect"?
    - Is there any visual feedback with explanations of what the user can see?
    - Is the feedback correctly given?
- Scoring
    - Is there a scoring of the user's performance after training units?
    - Can the user understand the formation of the scoring?
    - Is the scoring useful for the motivation and the planning of the next steps?
- Pedagogical structure
    - Does the user receive any information about her/his learning progress?
    - Are there aims formulated that can or should be achieved in a given time?
    - How much guidance and how much liberty exist when getting through the training programme?

3 Results

It reveals that most, and sometimes all, tools show substantial weak points on all aspects of pronunciation training. What follows are summaries of the collected comments.

3.1 Level of proficiency and individual needs

The level of proficiency has of course an important impact on the skills to be improved. For this reason it is astonishing to see that some of the inspected tools not even consider the level of proficiency. Most tools were ignoring learner characteristics like age, motivation and individual aims. However, L1 was often but not always taken into account at the beginning of the programme.

A popular way of test for level of proficiency is to administer placement tests at the beginning of the training programme. In some tools a self-assessment is also used to complement the placement test. Sometimes it was unclear how the placement test was incorporated, i.e.

whether it was really used to select exercises and to set aims for the individual user. For some tools the assumed level based on the placement test was not realistic. This could be easily seen by the exercises that were too easy to perform for intermediate and advanced levels. In one case, after the performance of the placement test the user was informed that s/he will receive the test results in 10 days (which has led to great amusement of the evaluators).

It was hard to see whether the exercises were tailored to the specific L1. Serious answers to such a question would need more time of evaluation.

3.2 Instructions

In many cases there was a lack in clarity of instruction. There were often exercises where the user did not know what to do. Although the majority of exercises seemed to be clear or clear enough in order to be performed, the number of unclear instructions was unexpectedly high.

This picture is continued when asking whether it was clear why a specific exercise should be performed. Frequently, it remained a riddle for the user what a specific task was good for.

There are huge differences regarding the orientation in the training programme. For some tools it was very clear which training units were to be done, in which order, how to proceed, and where to get an overview of the individual programme. In other tools, the user had the feeling to be lost, either completely or with much effort the orientation was newly acquired. Some tools can be expected that they are not adapted to the users' expectations.

3.3 Quality and types of the exercises

The quality of the exercises with regard to pronunciation was very often insufficient. In some CALL systems there were only a few exercises on pronunciation matters. In one system "pronunciation" was handled only as a matter of spelling. Dedicated CAPT tools of course concentrated on phonetic topics. However, this was partly done by considering single sounds – without any further phonetic context setting syllables, words and sentences aside.

Regarding the covered topics of pronunciation problems in L2 learning two phenomena were noticeable. First, in some tools important problems of pronunciation, like liaison in French, were ignored. Second, other tools provided an overflow of information, with the consequence that the user cannot distinguish what is more important and what is less important.

As in classical pronunciation instruction in classrooms, the correct acquisition of vowels and consonants is the widespread goal. Prosodic topics such as word stress, sentence accent and phrasing, variation of speech tempo, or expressive speaking styles were rarely covered. This is also valid for training and testing of speaking fluency.

Most of the tasks for training of speech production skills were based on ASR. The goodness of corrective feedback in repeating exercises seemed rather arbitrary. For one system even native speakers were rejected (for English as target language), in contrast to a simulated strong foreign accent. Sometimes only an unnaturally hyper-articulated speaking style led to good results. In addition, the correctness of the results of ASR-based exercises was often questionable. Sometimes the recognition of the actual spoken words was simply incorrect.

The inspected CAPT tools showed great differences regarding the diversity of exercises. Speaking exercises using ASR are widespread. In contrast, listening exercises seem to have secondary, if any, importance for most tools. However, one tool is mainly based on listening training but does not provide any exercises on speech production.

Exercises that make use of "traditional" material, e.g. using minimal pairs, are not very frequent. Speech perception skills are mainly considered as speech understanding that was trained when a semantic context was given. Explicit treatment of letter-to-sound correspondence was not very frequent.

3.4 Feedback

The feedback given during and after the completion of single exercises was either not existent or a simple "correct-incorrect". If there was visual feedback, e.g. in terms of curves, the content and the use of these displays were left unexplained, i.e. the relevance of the parameters displayed is questionable and often there is no link with any objective parameter accessible to the learner who thus gets lost.

In addition, the speech analysis algorithms behind the provided feedback are often not sufficiently robust because this domain remains an open challenge which does not receive appropriate attention. There is a strong difficulty with speech analysis parameters, i.e. formants for instance, and to some extent even with F0. Furthermore, feedback was not designed to incorporate confidence measures to adapt its form.

There was an infrequent use of the various forms of explicit (corrective) feedback such as explicit recast and explicit correction or those forms prompting a student-generated repair like clarification requests or a repetition as question. Metalinguistic feedback on a specific error of the user was also rarely observable.

Display is sometimes clumsy, or even very clumsy in some games which lead to the learner's death. This could be acceptable with teenagers accostumed to electronic games, but far less to other learners.

3.5 Scoring

Some of the tools do not show any scoring mechanism. As a consequence a control of the learning progress was reduced to the simple performance of the different tasks. The majority of the tools applied some scoring mechanisms. However, for most of them explanations were missing. The meaning and the logic of the scores kept hidden. In these cases a control of the learning progress had no real meaning. Nevertheless there were also examples where the user's performance was scored after a training unit in a reasonable way and which could also be used for the planning of next steps.

3.6 Pedagogical structure

Information about the learning progress while getting through the training programme was not always provided in the tools. General or individually agreed aims in which certain tasks should be achieved in a given time were rather rare.

Some of the programmes can only be passed through a given guidance whereas others leave it completely open to the learner which kind of topics and exercises should be selected. One problem with the guided forms is that sometimes the activities of the tasks were repeated in their type.

4 Discussion

The evaluation was performed by experts in phonetics and/or signal processing and ASR, not “real” users. This allows only a limited insight what we can expect in reality. The following

discussion aims at drawing some perspectives, recommendation and guidelines for each of the points assessed in the results section.

4.1 Level of proficiency and individual needs

To take the level of proficiency of the learner into account should be an obligation for a serious CAPT tool. Placement tests are a good instrument to find out the level of proficiency of the individual learner which could be easily combined with a self-assessment.

Astonishingly, no further individual needs were asked for such as age, motivation and personal aims. A 10-year old learner who would like to improve the L2 skills acquired in school should be getting other exercises and feedback than a 40-year old learner who has interests in touristic purposes. Likewise a university student with an interest in reducing the foreign accent should get a different learning scenario than somebody who just intends to master the most basic pronunciation problems known for his/her L1.

4.2 Instructions

Clarity of instruction is one of the prime principles in didactic work. It is amazing to see how often this principle was either ignored or not tested with real users. In the same vein, the missing orientation in the training process should be mentioned.

4.3 Quality and types of the exercises

The quality of the exercises can be improved in many cases. First, the exercises should cover the most important phenomena of pronunciation problems. In addition, the range of phenomena targeted by one exercise or feedback should not be too extended. Though spelling should not be excluded as a topic for pronunciation exercises, it cannot represent the base for the PT of vowels and consonants.

Prosodic topics got too less attention. Listening exercises with a focus on word stress and sentence accent could be easily generated. Automated fluency testing, also with read material, is still a feature which should be followed.

Despite a good progress in ASR-based tasks the correctness of the recognition result must be the ultimate goal. It would be a plus for the user if s/he could use the own recordings for further manipulation.

The inspected CAPT tools showed great differences regarding the diversity of exercises. Speaking exercises using ASR are widespread. In contrast, listening exercises seem to have second, if any, importance for most tools. However, one tool is mainly based on listening training but does not provide any exercises for speech production.

Exercises that make use of "traditional" phonetic material, e.g. using minimal pairs, could be easily used in listening and also speaking exercises. Generally, there is a deficit in phonetic listening exercises where sounds and prosodic patterns should be discriminated or identified. Ideally, listening exercises combine analytical listening on a phonological level with semantic listening that aims understanding in a context. Moreover, more explicit exercises on letter-to-sound correspondences would be desirable. Regarding speaking exercises reproductive forms (e.g. reading words or sentences, or picture naming) could be complemented with productive forms where speaking means more than repeating, for instance forming plural forms or finding antonyms.

4.4 Feedback

The frequently observed non-existence of feedback of any kind reveals one of the most important deficits in the inspected CAPT tools. But the mere existence of feedback does definitively not improve the situation. Missing explanations of graphical feedback made experts sometimes wish to reject those sources of misunderstanding and frustration.

Simple forms of feedback like "correct-incorrect" of course represent important information. But there is a huge potential of exploiting more explicit forms of feedback, particularly metalinguistic feedback which could be adopted to the needs of the individual user.

4.5 Scoring

The good examples of scoring show that it is possible to apply such a mechanism in a CAPT tool. The scoring results should of course be understandable, if not they should be avoided. Scoring is definitively a benefit but probably not an indispensable requirement for giving the user an instrument to follow the individual learning progress. Its precision should be consistent with the robustness of the diagnosis, i.e. rather roughly if robustness is low, and at a finer level if the robustness is at a higher degree.

4.6 Pedagogical structure

A serious learning tool should also provide the possibility to display the aims of the learning procedure. These aims can be general aims for "everybody" or, more preferably, can be adapted to the individual needs of the learner.

In strongly guided programmes a change of activities would lead to more "fun of learning", so as to keep the user engaged in the learning activity. In unguided forms at least suggestions for having routes to get through an individual programme would be helpful.

5 Conclusions

Although the applied assessment must still be regarded as informal, and although the selected CAPT tools might not be an optimal sample for representing the state of the art, the results clearly show that there is a lot to improve regarding the clarity of instruction, the quality of exercises, the clarity and appropriateness of scoring, the diversity of feedback methods, the presumed benefit for various types of users as well as the usage of ASR. Despite various good approaches regarding graphics and game-like exercises there are obviously missing links between the pedagogical expertise in phonetic training on the one hand, and software development including usability engineering on the other hand. An important task for future studies will be to develop more formal procedures of how to assess CAPT tools, also by considering existing approaches [1,2,4,5].

As depicted in Figure 1 the technical knowledge is at the core of each CAPT tool. There are various levels for this knowledge, be it ASR for non-native speech or graphical solutions or content management, and other levels, which are all independent of each other at the first place. More input from other sources of knowledge is needed at many places: from language learning and teaching practice (pedagogical knowledge) and from phonetics and phonology, particularly when specialised in non-native speech research. Particularly the knowledge emerged from the interface between the research-oriented disciplines of phonetics and phonology, and the pedagogical practise is important here. In addition, the testing with users in real learning situations is an important factor which seems to have spared out quite

frequently. It is essential that the results of usability testing are integrated in the technical circle to improve the robustness of the tool.

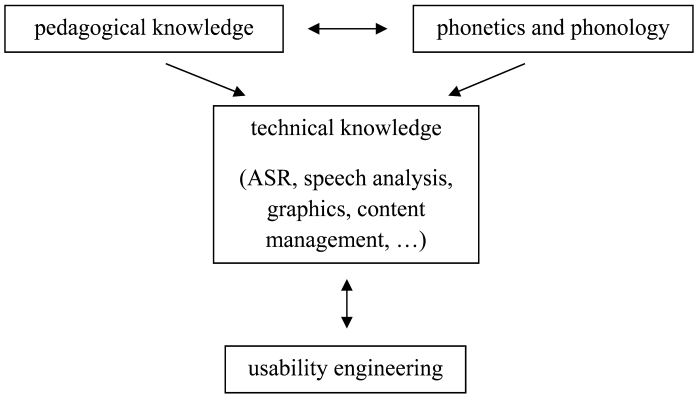

Figure 1 – Sources of knowledge needed for CAPT tools.

6 Acknowledgements

The authors thank the participants of the workshop on "Feedback in Pronunciation Training" who joined us in evaluating some CAPT tools (as a pilot test). Moreover, our thanks go to all colleagues in Nancy and Saarbrücken who were actively involved in testing, particularly Lennart Schmeling.

Literature

[1] CHAPELLE, C. 2001. Computer Applications in Second Language Acquisition. Cambridge: Cambridge University Press.
[2] CUCCHIARINI, C., NERI, A. & STRIK, H. 2009. Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. Speech Communication 51, pp. 853-863.
[3] DIELING, H. & HIRSCHFELD, U. 2000. Phonetik lehren und lernen. München: Langenscheidt.
[4] HARDISON, D. 2004. Generalization of computer-assisted prosody training: quantitative and qualitative findings. Language Learning and Technology 8 (1), pp. 34-52.
[5] HINCKS, R. 2003. Speech technologies for pronunciation feedback and evaluation. ReCALL 15 (1), pp. 3-20.