

Computer Assisted Pronunciation Training: The four 'K's of feedback

Thomas K. Hansen

NISLAB, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark.

One of the biggest problem areas in relation to Computer Assisted Pronunciation Training (CAPT) applications, designed for Second Language Learning (SLL), has been that of feedback. As with human teachers, the ultimate goal of CAPT is to provide instantaneous, individualized feedback to the user on the overall quality of the pronunciation made. Attempts at accomplishing this goal have seen the light of day in many different ways. Some applications provide intonation curves and some provide spectrograms, color-coding or numerical scores. Unfortunately these feedback modes rarely inform the user of what is wrong instead relying on the users to self-correct through trial and error [2].

In this paper, I suggest that feedback should adhere to the four 'K's, being: 1) Quantitative 2) Qualitative 3) Comprehensible and 4) Corrective.

I further try to illustrate the four 'K's by relating them to the TASK-LT program which is a CAPT application currently under development at the Natural Interactive Systems laboratory.

Keywords Computer Assisted Pronunciation Training; learning feedback.

1. Introduction

Teachers have always done it. At the turn of the 20th century wax records started doing it. Radio broadcasts did it and then TV started doing it. In the 60s language laboratories started doing it. The Walkman and car cassette player did it too and it carried over into the portable CD. The 60s saw it in a small measure and, finally, in the 80s, the personal computer started doing it on a large scale.¹ Automated language teaching and learning has indeed been around and on its way for a long time.

When perusing bookshops and browsing the homepages of software developers today, it is quickly revealed that a myriad of different language learning programs focusing on various languages are available for the enthusiastic learner. These take the form of popular commercial products such as *The Rosetta Stone* or *Talk To Me*, as well as several prototypes described by various research institutes around the world.²

The potential advantages of CALL, and maybe in particular CAPT applications, have often been debated and are nowadays well-known [2]. The sometimes over-crowded classrooms which often leave little time for teachers to focus on individual problems, especially as they relate to the learner's pronunciation ability, are but one of the factors. Others revolve around the possibilities of providing private, stress-free environments, self-paced learning and unlimited patience from the artificial tutor.

But CALL/CAPT applications, as well as the research area on the whole, have also been heavily criticized and perhaps rightfully so. The core of the criticism concerning actual applications centers primarily on whether these are guided by sound pedagogical rules or simply by the possibility of providing fancy looking multimedia effects [3]. Other critical voices emphasize the seeming lack of a standard for CAPT development or a missing red thread [4] [5]. But the most prominent topic of discussion in recent years, relates to the issue of feedback within the applications.

¹ See www.history-of-call.org

² See for instance <http://www.nis.sdu.dk/projects/CAPT/> or <http://www.speech.kth.se/ville/>

2. The issue of feedback: pre- and post ASR technology

The last decade has spawned a series of CAPT applications that employ Automatic Speech Recognition (ASR) as part of the design and feedback strategy. With the development and integration of ASR technology, the CAPT area seems for the first time within reach of being able to provide instantaneous, individualized feedback to the learner.

Prior to the inclusion of ASR, as noted by Hincks [1] and Neri [2], many applications relied on the use of signal analysis software which enabled visualization of the speech signal through either spectrograms or waveforms. The learner would then be able to compare his or her speech input with an already stored model of the target word, as pronounced by a native speaker. But as also pointed out by Neri [2] these feedback modes were not originally designed with the learner in mind, but rather to be used by professional phoneticians. Hence, often it was required that the learner receive assistance from a trained professional for interpretation. And even though the use of signal analysis software has been found to yield helpful results in terms of providing intonation curves for learners of tone languages or simply prosodic training [1], they really do not inform the learner of possible segmental errors.

Other applications asked the learner to pronounce a set of words that were recorded and then either had to be uploaded to a server or sent directly to a teacher for evaluation. Although this mode of practice ensured an individualized mode of practice and feedback, it proved to be both time consuming and to go against the idea of instantaneous feedback. This left the user free to continue making the same mistakes over and over rather than having them corrected at once.

Even after the inclusion of ASR, feedback strategies were and continue to be under strong criticism in one way or another.

Neri [2] notes that some applications make use of a strategy by which a mispronounced word is highlighted in a different color in order to enable the user to identify where the mistake was made. Or they include a word- or nativeness score ranking the user depending on some predefined values.

The main problem here lies in the fact that although colors and scores can give the learner an indication of where a mistake is made, or how well he or she is doing, it cannot indicate what the mistake was or how to correct it. The only choice for the user is to attempt a better result through trial and error. Hincks [1] notes the same problem in other applications given the probabilistic nature of ASR technology. It becomes possible to provide the learner with a numerical score as an expression of how well the word or sentence was recognized, but again the user is left to trial and error correction.

However, color coding or numerical scores do still seem to have value in CAPT, but as a competition element rather than in terms of constructive feedback. Receiving a numerical score on a pronunciation attempt could easily inspire the learner to try and receive a higher score on the next attempt, not unlike a grading system or a computer game.

The ISLE project, as described in [2] was a commendable attempt at using ASR in a CAPT application. Here the specific error words were highlighted for the user to click. Subsequently the user was informed that vowel x should be pronounced like the one in word y and not the one in word z.

Although this seems to be exactly the kind of feedback which is being sought after, it also revealed the perhaps most fragile aspect of ASR, namely erroneous feedback. The application simply tended to return too much erroneous feedback to be of value to the user and the project was abandoned.

There can be no doubt that integrating ASR in CAPT is by far the most valuable component in the search for instantaneous, individualized feedback to date. But it is also painfully clear that there are still many short-comings.

3. Short and long term goals

The ultimate goal for an ASR based CAPT application is to emulate a real-life teacher. A competent well-qualified language teacher is able to correct a learner's pronunciation on single-sound, word or sentence level. The same teacher is also able to correct the same learner's segmental, prosodic and even pragmatic mistakes. A CAPT application is currently not.

In 1996 Hubbard [7] stated that the question is no longer whether computers can help, but how? So perhaps research should be divided into long and short term goals? The short term goals are what we can do with the present level of available technology, and the long term take into consideration our goals for the future? It seems a fair assumption that much of the functionality incorporated in CAPT today, such as ASR, animated agents, speech synthesis, are still technologies in their early stages of development. But in spite of this many of them have proven helpful to learners.

4. TASK-LT and the four 'K's of feedback

TASK-LT, as seen in Figure 1, is a CAPT application currently under development at the Natural Interactive Systems Laboratory as part of a PhD project. The aim of the application is to, initially, teach American English single-word pronunciation to Danish learners. The application itself targets both perception and production in the learner as well as attempts to provide motivational, competitive and educational aspects.

Much of the language material used in TASK-LT is based on teaching materials developed by the late John M. Dienhart who taught the course American Phonetics at the University for many years, and has therefore been used in language laboratories for decades.

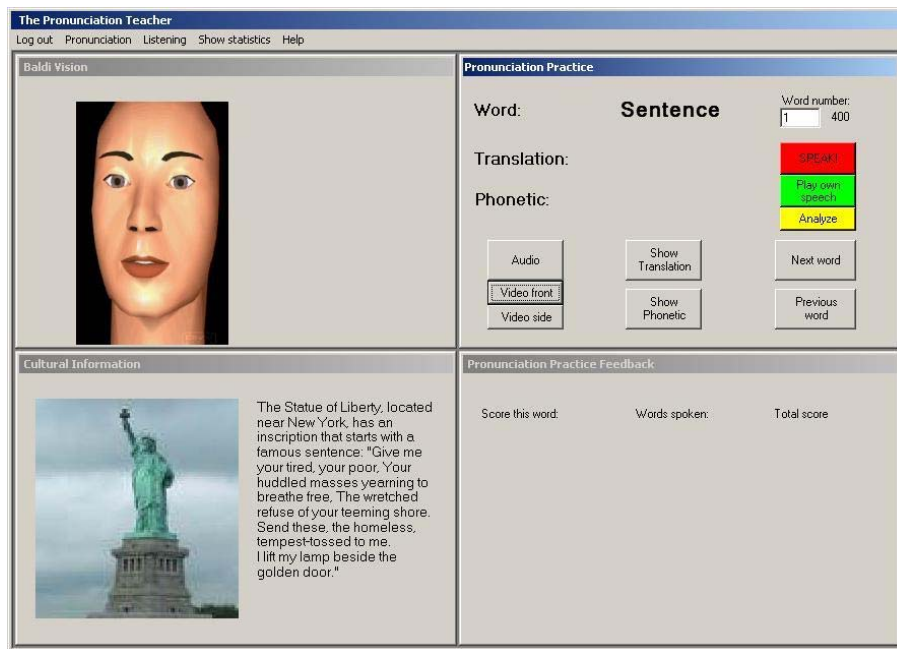


Fig. 1 The TASK-LT application as seen in Pronunciation Practice Mode

TASK-LT employs an ASR engine developed by MULTITEL in Belgium and was trained on the TIMIT corpus for phoneme recognition.

The development philosophy underlying the application centers on the four 'K's of feedback. And although not all of them are implemented as yet, the capabilities are present within the system. The lack of implementation is due to an incremental build, hitherto unanswered questions, as well as a current lack of technological capability.

The four 'K's, are a phonetic play on words and stand for *Comprehensible, Qualitative, Quantitative and Corrective*.

Comprehensible: as touched upon in section 2, a learner needs to be able to understand the kind of feedback which is returned by the application without having prior knowledge of phonetics, phonology or any other kind of language training. Therefore the application's primary means of feedback is textual, attempting to explain to the learner how the mispronounced segment should be said properly. There are numerical scores and color-coding in TASK-LT as well, but these serve other purposes. The color-coding is semiotic in nature, reminiscent of a traffic light with orange as an added color. The purpose is to point out to the learner how grave the mistake is. Red means that the recognition was very low and green means very high, as seen in Figure 2. The numerical score is intended to provide the application with a competitive element, urging the learner to try and receive an even higher score on subsequent pronunciation attempts. It should be stated that presently the score settings are fairly arbitrary in nature and that 'natural' settings are being explored. In Figure 2, the <e> is colored yellow and the <t> is colored red.

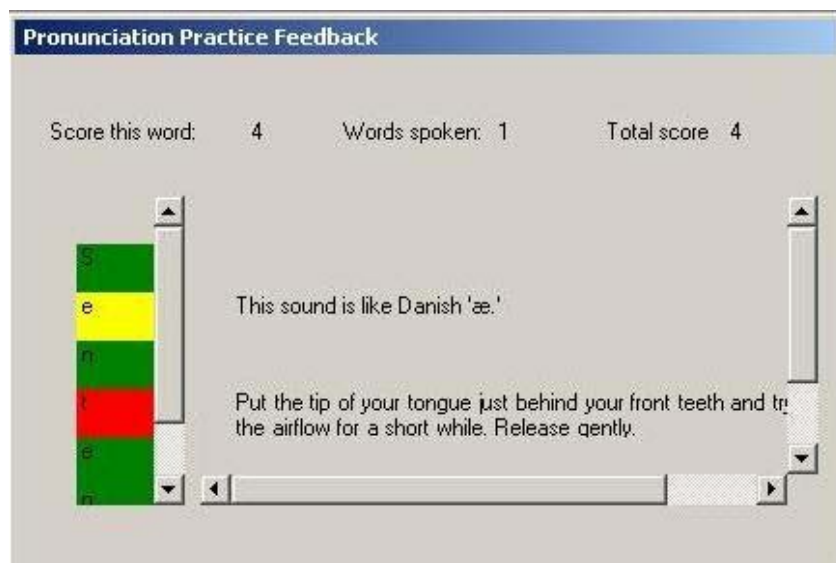


Fig. 2 Segmental feedback in TASK-LT

Qualitative: The application must be able to determine whether the correct vowel or consonant was used in the pronunciation and if not, make the user aware of the mistake. Obviously there is a difference between saying mice and lice and the learner's attention should be brought to this.

Quantitative: In the sense that the length of certain segments should be evaluated. Although the problem is not overwhelming in American English, there are languages where the length of a segment functions as a distinctive marker. In English we find examples such as <pig> and <pick>, where the length distinction is due to pre-fortis clipping, in which vowels tend to be shortened in front of unvoiced stops. Estonian [7], for instance, is a language where a short, long and overlong contrasts can be found in both vowels and consonants. In case of a learner immersed in Estonian, these differences should be pointed out., or the learner may find him or herself training the wrong word. The quantity measure is as yet not implemented in TASK-LT although the possibility is there. The question that remains to be answered is what would be appropriate settings for pointing out length contrasts.

Corrective: As pointed out by Neri [8] the ideal phase of a system should be able to diagnose the error which has occurred and suggest the appropriate remedial steps. In essence, this was the attempt which was made in the previously described ISLE project. But as also pointed out by Neri [8] ASR is as yet not able to perform this function due to the immaturity of the technology. The approach taken in TASK-LT is to guide the learner toward a better pronunciation by highlighting where the mistake(s) was made and offering textual advice on how that particular segment should be pronounced. The segmental feedback

here is based on knowledge of phonetics and how segments are pronounced depending on their place of occurrence in the target word. Hence the description aims at either drawing parallels to the native language or by giving advice on how to pronounce the segment in its current context, as illustrated in Figure 2. Furthermore the learner can press the segment in question and hear the target pronunciation.

4. Conclusion

There are feedback issues that have been left out of the above description such as prosody, rate of speech and can ASR confidence scores be used. Although these are important issues to solve, I have concentrated on the two areas which seem most heavily debated, namely comprehensibility and corrective feedback. In addition I have offered a point of view on how segmental feedback can be addressed and also tried to draw attention to the fact that certain aspects as they relate to ASR and feedback simply do not seem possible given the current level of technology. Nevertheless it is an undisputable fact that applications are flooding the market at the moment and it seems prudent to focus on the *can do's* and not just *can't do's*.

Acknowledgements To Per Thue Olsen for his invaluable assistance in creating the TASK-LT software. Christophe Ris at MULTITEL for all his advice and help in setting up the ASR system and Dominic Massaro and Michael Cohen from the University of California, Santa Cruz for providing the Baldi segments.

References

- [1] Hincks, Rebecca, Speech Technologies for pronunciation feedback and Evaluation, ReCALL 15(1) 3-20 (2003).
- [2] Neri, Ambra, Cucchiarini, C. Strik, H. And Boves. L. The pedagogy-technology interface in Computer Assisted Pronunciation Training. Computer Assisted Language Learning, 15:5, pp. 441-467 (2002)
- [3] Neri, A. Cucchiarini, C. Strik, H. Feedback in Computer Assisted Pronunciation Training: When technology meets pedagogy. Proceedings of CALL Conference "CALL professionals and the future of CALL research." Pp. 179-188 (2002b)
- [4] Chapelle C. Computer Applications in Second Language Acquisition. CUP 2001
- [5] Levy, M. Computer-Assisted Language Learning. Context and Conceptualization. Clarendon Press, 1997.
- [6] ed. Pennington, Martha C. The Power of CALL Athelstan,(1996).
- [7] Wik, Preben. Hands-on Instructions to the VISPP summer school, Palmse Estonia August 2005. available at <http://www.speech.kth.se/ville/publications/VISPP2005.pdf>
- [8] Neri, A. Cucchiarini, C. Strik, H. Automatic Speech Recognition for second language learning: How and why it actually works. Proceedings of 15th ICPHS, Barcelona, Spain, pp. 1157-1160 (2003)

...