Wizard-of-Oz Test of ARTUR - a Computer-Based Speech Training System with Articulation Correction

Olle Bälter[1], Olov Engwall[2], Anne-Marie Öster[2] and Hedvig Kjellström[1]

[1]Interaction and Presentation Laboratory
Nada, KTH (Royal Institute of Technology)
SE-100 44 Stockholm, Sweden
+46 8 790 6341
{balter, hedvig}@kth.se

[2]Centre for Speech Technology (CTT)
TMH, KTH (Royal Institute of Technology)
SE-100 44 Stockholm, Sweden
+46 8 790 7565
{olov, annemarie}@speech.kth.se

ABSTRACT

This study has been performed in order to test the human-machine interface of a computer-based speech training aid named ARTUR with the main feature that it can give suggestions on how to improve articulation. Two user groups were involved: three children aged 9-14 with extensive experience of speech training, and three children aged 6. All children had general language disorders.

The study indicates that the present interface is usable without prior training or instructions, even for the younger children, although it needs some improvement to fit illiterate children. The granularity of the mesh that classifies mispronunciations was satisfactory, but can be developed further.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces - *Auditory (non-speech) feedback, Prototyping, User-centered design*

General Terms

Design, Human Factors

Keywords

Computer-based speech training system, Wizard-of-Oz, user interface.

1. INTRODUCTION

A hearing impairment may lead to unintelligible speech caused by difficulties hearing what is wrong with the own speech production. Speech therapists can help, often resulting in dramatic improvements. One of several important tools for these therapists are computer-based speech training (CBST) systems.

There is a wide range CBST systems used in speech training for children with hearing and/or speech impairment. Examples are SpeechViewer [2], Box of Tricks [23], Indiana Speech Training Aid (ISTRA) [24], Speech Illumina Mentor (SIM) [22], Speech Training, Assessment, and Remediation system (STAR) [5], and the OLP-method [17].

Of these, both SpeechViewer and Box of Tricks are extensively used and acknowledged, as the CBST therapy has shown to be very efficient, especially in the instruction phase of speech training [23, 18]. Research has however shown [11] that there is a need of CBST systems that can support the learner without the presence of a speech therapist. A major drawback of most CBST systems today is their need for support by a trained specialist. Motor learning theory in speech development postulates that repeated practice with accurate feedback is essential to establish automaticity and to transfer skills to untrained situations [26]. This is the most important element in a speech therapy program but the most difficult for a therapist to carry out, due to time limits. The target production must be repeated and practiced in a variety of contexts until the articulation can be made without deliberate planning. To use computer-assisted speech training in this situation may be particularly helpful to motivate the child to significant amounts of additional training [19]. Children who are born with a severe auditory deficit have a limited acoustic speech target to imitate and compare with their own production. Other senses must replace the auditory feedback that hearing children use when they learn to speak. In general, CBSTs do this by offering more or (often) less advanced visualization of the *acoustic* signal as feedback. For a hearing-impaired child with limited notion of the acoustic targets it is however often more fruitful to focus on visual or tactile properties of the pronunciation. The virtual teacher Baldi [15] provides audiovisual instructions on how to produce the training sound correctly on the articulatory level, but without relating it to the student's own production. As both imitation and self-correction are important factors in speech learning, we believe that it is of primary interest to be able to show not only correct articulations, but also how the student should alter his/her production to reach this target.

1.1 ARTUR – the ARticulation TUtoR

For that reason a new CBST system, the ARticulation TUtoR (ARTUR) [9], is presently being developed at KTH (Royal Institute of Technology), Sweden. The goal of ARTUR is a speech training aid, with a virtual speech tutor Artur (which is the Swedish version of the name Arthur), who can use three-dimensional animations of the face and internal parts of the mouth (tongue, palate, jaw etc) to give feedback on the difference between the user's deviation and a correct pronunciation.

The main feature of ARTUR is the ability to give clear instructions on how to improve the articulation as feedback and illustrate salient parts of the instructions. For example, if a user practicing the r-l distinction pronounces "*Harry Potter*" as "*Hally Pottel*", Artur would reply e.g.: "*That sounded more like Hally*

ASSETS'05, October 9-12, 2005, Baltimore, Maryland, USA.
Copyright 2005 ACM 1-59593-159-7/05/0010...$5.00.

Pottel. Try to retract the tongue tip and make the contact between the tongue and the palate with the edges, instead of the middle, to get a vibration of the tongue tip".

The use of a talking head with internal parts is a key feature, as phonetic features that are hidden in a human speaker can be displayed. The perception of speech through lip-reading is difficult because many articulatory and acoustic features of speech are not easily accessible from visual observation. Acoustically each speech sound is unique, but visually many sounds are difficult or impossible to discriminate from a view of the speaker's face, as they have almost identical visual articulatory movements or invisible articulation [10, 14]. With a talking head, on the other hand, parts of the anatomy may be removed to display the manner and place of articulation (c.f. Figure 1 for an example).

A main focus group of the project is hearing-impaired children with residual hearing, who can benefit from the audiovisual feedback in the speech-training program. As acoustic and visual speech are complementary modalities, learning will be more robust and efficient with multimodal training than with either modality alone.

ARTUR involves several subtasks, shown in Figure 2, of which the majority are still to be implemented:

*Audio-visual detection of mispronounced speech***.** The input to the system is the user's utterances and the aim is to detect deviations between the target and the user's pronunciation, based mainly on acoustic data. This is a non-trivial extension to speech recognition, as large mispronunciations may occur. On the other hand, the expected input from the user is generally known (the exercises consist of repeating words or sentences or practicing a specific articulation) and can be compared to a target utterance using forced alignment. One method to improve the speech recognition is to add visual information [16] as correlations between e.g. jaw and lip position and speech acoustics can be exploited [3]. Facial data will hence be used to increase the robustness of the mispronunciation detection.

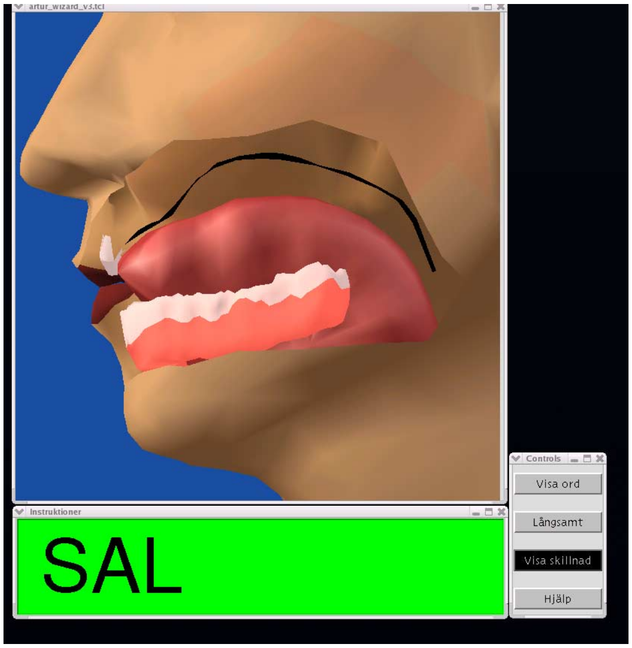

Figure 1. The user interface for articulatory feedback.
Top left: side view of the talking head model, with a part of the chin removed to make the intra-oral articulation visible. Bottom left: training word ("SAL", /sɑːl/, meaning "hall, room, ward" in Swedish), with the green color indicating that the student may speak. Right: user control buttons (refer to the text for details).

Marker-less tracking of facial features from video: The facial data, such as jaw position and mouth opening, is extracted from video images of the face. This can be done either by fitting a three-dimensional model of the face to the face in the video images [1, 13] or by training two-dimensional face appearance models from a large database of face images [6].

*Articulatory inversion***:** The next step is to recreate the user's motion of the face and vocal tract from the speech signal and face parameters. The visual input is important as there is a many-to-one mapping of acoustics to articulation, which means that the articulation cannot be recovered from the speech signal alone. As there is a significant correlation between the face and the tongue positions, facial data is used to improve the articulatory inversion [8].

*Articulatory model***:** The user's and the correct articulations are synthesized using the models of the face [4] and vocal tract [7] developed at KTH. The vocal tract model is articulatorily correct as the tongue shape is based on statistical analysis of Magnetic Resonance Imaging (MRI) data and the articulatory movements are modeled from Electromagnetic Articulography (EMA) measurements, both for the same subject.

Adaptation of the model to the user. The shape of the face and vocal tract varies between individuals, and the articulatory inversion requires that the model is adapted automatically to each new user. This will be done based on acoustic input and initial information on the speaker's age.

*Feedback display***.** The output of the system, an articulatory representation of the training utterance, requires much attention. It is crucial that the feedback is comprehensible, useful and motivating for the student. The current Wizard of Oz study was hence carried out as a first step to test and refine the human-computer interface and feedback display.

The Wizard of Oz study was made before spending time on developing the speech technology components, as the functionality of the interface will influence the requirements on the components. The study further served the purpose of collecting audio and video data that will be important training material for the mispronunciation detection.

2. METHOD

2.1 The Wizard of Oz set-up

The set-up of the Wizard of Oz system differed from the automatic system in the aspects shown in Figure 3. The mispronunciation detection and the articulatory inversion were performed by a phonetically trained human Wizard (the second author of this paper), some system tasks were disabled and the audio and video recordings were stored to create an audio-visual database.

The user interface, shown in Figure 1, consisted of one window displaying the virtual tutor Artur (implemented as a virtual face of

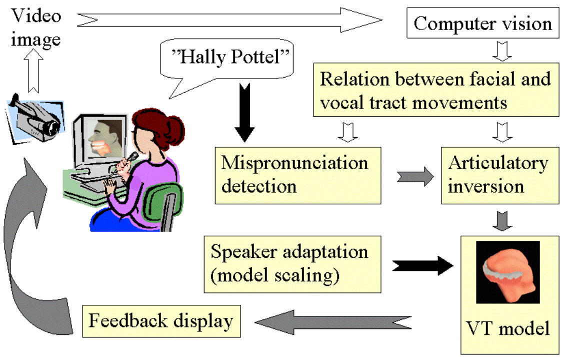
Video image
Computer vision
"Hally Pottel"
Relation between facial and vocal tract movements
Mispronunciation detection
Articulatory inversion
Speaker adaptation (model scaling)
Feedback display
VT model

Figure 2. Schematic overview of the ARTUR system. Black arrows indicate actions performed on the acoustic input, white on visual and grey on the combined audiovisual.

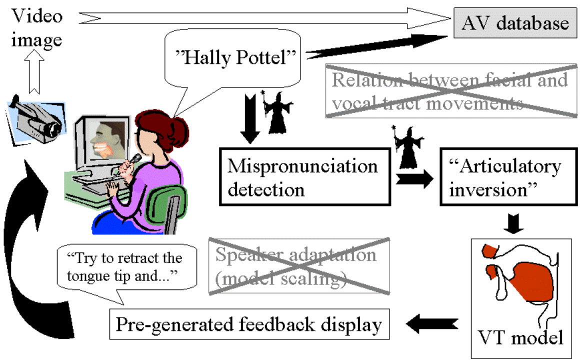
Video image
AV database
"Hally Pottel"
~~Relation between facial and vocal tract movements~~
Mispronunciation detection
"Articulatory inversion"
"Try to retract the tongue tip and..."
~~Speaker adaptation (model scaling)~~
Pre-generated feedback display
VT model

Figure 3. Schematic overview of the Wizard of Oz version of the ARTUR system. Black wizards indicate tasks where the wizard replaced the automatic system.

an approximately ten year old boy) and his articulatory feedback images, one text window showing the training words and subtitling of all Artur's utterances (as an additional support for hearing-impaired users) and one set of interaction buttons.

Each test began with Artur introducing himself and explaining the training procedure. Artur uses pre-recorded natural speech and time-aligned articulation movements generated from a text-to-visual-speech synthesizer [4, 7]. During the introduction, the student was given the possibility to test the interaction buttons, see Figure 1: "Show word" (Visa ord), "Slow" (Långsamt) and "Show difference" (Visa skillnad). Pressing "Show word" resulted in a repetition of the animation of the training word articulation; "Slow" in a slow-motion display of the articulation and "Show difference" in a still picture showing the correct and the student's articulation with the most important difference highlighted by green (correct articulatory feature) and red (incorrect) circles. The fourth button "Help" repeated the explanations given in the introduction.

The session consisted in repeating 18 words after the tutor. The words were 9 minimal pairs of one- or two-syllabic nouns or verbs starting with one of the fricatives (using IPA [12] notation) /s/ or /ɧ/ (voiceless velar fricative with rounded lips) preceding the vowels /ɑː, eː, iː, uː, ʉː, yː, oː, ɛː, øː/ in the Swedish words 'sal' *vs.* 'sjal' (ward *vs.* scarf), 'se' *vs.* 'ske' (see vs. happen), 'sol' *vs.* 'kjol' (sun *vs.* skirt), 'sula' *vs.* 'skjul' (sole vs. shed), 'sylt' *vs.* 'skylt' (jam *vs.* sign), 'säl' *vs.* 'skäl' (seal *vs.* reason), 'söta' *vs.* 'sköta' (to sweeten *vs.* to nurse). The training began with the word starting with /s/ for each pair.

During the training session the student was placed alone in front of Artur in a sound-proof room and a microphone was fitted on the collar of the subject's sweater, see Figure 4. The Wizard of Oz system was run on one single computer, using a screen splitter to display the user interface on both the user's and the Wizard's screens. During the training session, the inputs from the user were vocal (uttering the training words) or with the mouse, whereas the Wizard controlled the feedback and encouragements using a cordless keyboard.

Outside the room the Wizard monitored the system and selected the appropriate feedback (see Figure 5), based on the user's acoustic utterances. The Wizard could choose from the ten different feedback options in Table 1, three encouragement utterances and three options to navigate between the training words (previous word, repeat the current or jump to the next). The feedback options were based on the assumed position of the tongue, which could be judged to be incorrect in place or/and

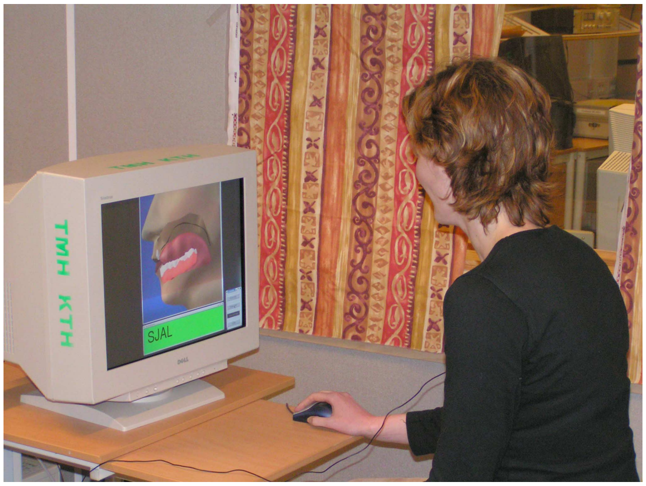

Figure 4. Test person in front of Artur. The window to the Wizard is visible in the background.

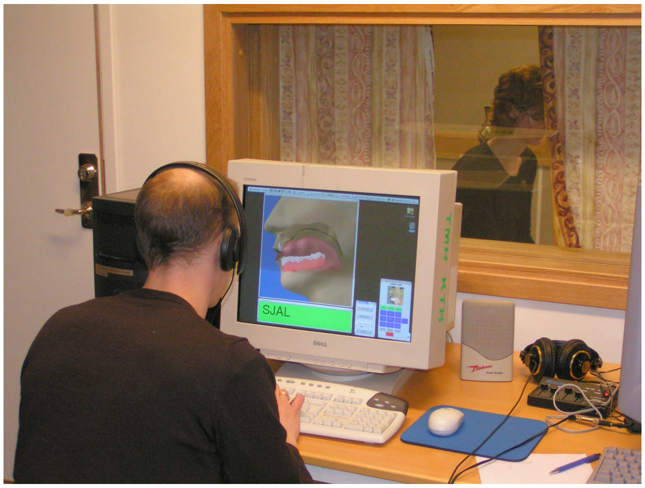

Figure 5. Wizard in front of the ARTUR control system. The test person is visible through the window.

Table 1. The feedback options available to the Wizard, complemented with descriptions of the most salient error.

	← Tongue position →			
← Tongue height →	Dental stop /t/; too constricted	Retro-flex stop /ʈ/; too constricted	Velar stop /k/; too constricted	
	/s/	Palatal voiceless fricative /ɕ/.	/ɧ/	Pharyngeal fricative /□/; too backward.
	Lisp	No audible fricative.	Fricative made between tongue edges and the teeth.	

manner of articulation. Note that the generated feedback depended on the training word, as, e.g., the detection of a word-initial /ɕ/ should result in a feedback indicating that the articulation should be more forward if the training word began with /s/, but more retracted if it started with /ɧ/.

Each feedback was of the type: 1) Initial encouragement + 2) The detected acoustic output (a word with the same word stem as the training word, but starting with the phoneme that the speaker made) + 3) Instructions on how to change the articulation; e.g., for the training word “sal”: *“Almost! Now you said ‘tal’. Try to lower the tongue tip, so that the air can pass.”*

2.2 Interviews

After the test, the student and the interface researcher (first author of this paper) went to another room separate from the test laboratory for an interview about the interface to ARTUR. The separate interview room was a measure to avoid effects from having the tested system present during the interview [20]. The interviews were semi-structured [21] using an interview script with open-ended questions, but with the possibility to probe the interviewee further if needed. It is especially suitable for interviews with children, where the interviewer has the possibility to explain and clarify if the child does not understand.

At the same time, the Wizard was debriefed about the training session in a more informal discussion with the remaining project members.

A test session lasted approximately ten minutes and the following interviews another fifteen minutes.

3. TEST SUBJECTS

Two user groups were tested: three children aged 9-14 with extensive experience of speech training and CBST systems, and three children aged six in the beginning of their speech training, with limited or no experience of CBST systems.

As a pre-study, a fluent second language learner was recruited in order to perform basic tests of the system, the instructions during the training session, and the interview script. This subject has Persian as his mother tongue, but is fluent in English and Swedish, and thus has experience of second language learning as well as CBSTs.

None of three older children (9-14) had any hearing difficulties, but all had language disorders. Classified according to ICD 10 [25] they all had a mixture of F80.1 ABC (expressive language disorder) and F80.2 ABC (impressive language disorder). At the time of the study these disorders had been dramatically reduced, but to a varying degree. One of them could speak practically without any difficulties; the other two were occasionally incomprehensible. All three followed the instructions from Artur without any assistance, but during the interview an adult accompanying the child assisted when the answers (or questions) were unclear. The accompanying adult was one parent, one speech therapist, and one teacher, respectively.

The group of three younger children (all six years old) all had several years of experience of speech training but little experience of CBSTs. None had any hearing difficulties, but all had language disorders, classified according to ICD 10 as F80.2B (general language disorder). At the time of the study, all three could answer yes-or-no-questions, but had limited abilities to describe things. These children had their speech therapist sitting next to them during the test. This was mainly to support the children during the Artur instructions. For practical reasons, the therapists stayed with the children during the entire test, but were quiet during the training session. The speech therapist then accompanied the child to the following interview.

Due to the involvement of children, and the difficulties of interviewing children, the children were prepared for the study by a visit by the interface researcher. There are several reasons for this:

1) To be able to make the purpose of the test clear for the child (that it was the system that was under scrutiny, not the child).
2) To make the child more relaxed for the test and interview by first meeting the interviewer in an environment that was familiar to the child.
3) To make the interviewer a familiar person for the child.
4) To make the interviewer and the Wizard aware of the child’s strengths and weaknesses before the interview.

One of these meetings took place in the home of the child, the other at the child’s school.

For the group of younger children, the interview script was adapted in order to fit the age group better. This was accomplished by replacing some words with simpler versions (e.g. “imitate” was replaced by “do the same”), and by making it possible to express opinions by pointing at iconic faces, see Figure 6. For comparisons of ARTUR to other CBSTs, paper slips were prepared with text and iconic pictures representing the different systems. These slips were given to the child in order to sort them after there liking.

A screen shot of ARTUR showing a side view of a head with tongue, teeth, jaw and palate visible (see Figure 1) was left at the school for the younger children a week before the test. The fact that this could have an impact on the test was discussed with the speech therapists. However, in this discussion, we came to the conclusion that the reason that these children had not seen a see-through picture similar to Figure 1, was not that it would be unnatural at this stage of training, but simply the lack of such pictures. The speech therapists said that if they had had a picture

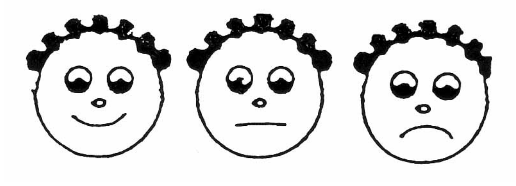

Figure 6. Iconic faces used for expressing opinions.

like that before, they would have used it. Besides this picture no information or instructions about ARTUR was given before the test.

4. RESULTS

The results of the study are presented below followed by a summary of conclusions based on the results in each section.

4.1 Adult second language learner

The adult testing the system was not originally planned to give any input to the design, but one if his comments afterwards was worth noting: "*It should be possible to practice pronunciation a few times before being evaluated by the system"*

A CBSTs used by a normal hearing person should not assume that feedback is necessary after each attempt. Learning a language with speech sounds that deviates greatly from the mother tongue, the student may hear that the speech production is wrong, and needs several attempts to get it right. This could also be a part of the process to make the student aware of the differences. The technical solution could be a button to abort the feedback.

4.2 Children aged 9-14

All three children were very positive to ARTUR. They described ARTUR as "*very intelligent and good and so*". The best part was the correction (the instruction on how to improve the pronunciation) on how to move the tongue more forward or backward. The main disadvantage was the limited number of sounds that could be practiced. The animation of the speech organs varied in popularity. One of the children said "really good"; one complained of technical flaws (e.g. ARTUR did not pronounce the entire word the second time); while the third was not that impressed, but thought that it was easy to understand.

None had any problems interpreting the feedback picture (see Figure 1) with the exception of the black line representing the hard palate that no-one could understand. One of the children described it as "it looks like a small secret passage" and wondered "where on earth is it located, maybe it is the nose and there is the air coming?"

All thought that imitating the animation worked well. One child thought that the instructions (voice and sub-titling) were better than the animation. Another mentioned that it was difficult to imitate the movements in the more backwards parts of the tongue.

All understood the function of the four buttons "See again", "Slow", "Show difference" and "Help", see Figure 1. However, few of them used them during the training session. When asked, they explained the reason for this was that they did not have any major problems pronouncing the fricatives, and after failing once they could get it right at the next attempt. One child thought that the "Slow"-button would be more useful for long words, such as "elephant", whereas no word in this test had more than two syllables. The children who did use the buttons did so on the second run of the training words, when they were more familiar with the training situation and wanted to explore the system.

When comparing ARTUR with other CBSTs, all found ARTUR better. The main reason was the feedback (correction) on the pronunciation. One child said "twice as good as SpeechView and Box-of-Tricks". There were however features of the other CBSTs that were better in those systems, such as the possibilities to practice on more varying sounds and scoring (getting points for correct pronunciation).

When comparing ARTUR to practicing the fricatives with their speech therapist, all considered ARTUR to be better (even though the speech therapist was present during this interview!). One child explained this with "It is nice to be able to practice on your own. It is relaxed." The same child also said that practicing with ARTUR felt "mysterious, strange" compared to practicing with his speech therapist. The explanation was that he had found new ways to move his tongue during the ARTUR session.

4.2.1 Conclusions

The idea of correction the pronunciation in ARTUR seems fruitful, especially the written/oral feedback. When it comes to the usage of the animation the results are mixed. This could be a learning effect. With more practice to interpret and mimic the animation the results may be more encouraging. We however believe that the animation speed of the articulatory feedback needs to be altered to separate the articulation that is practiced from the rest of the word. The animation now shows a slow, but natural production of the whole training word. As the children did not use the "Slow" and "Show difference" buttons, a better alternative may be to automatically show the part of the word that the feedback is focused on slower and exaggerated while the remaining parts are shown at normal speed.

The drawing of the hard palate clearly needs improvement.

The lack of use of the functions activated by buttons may also be caused by the novelty of these functions, but one explanation may also be that these children had too small problems with pronunciation.

More game-like features would increase the interest from the children to practice with the system.

4.3 Children aged six

These children had their speech therapists present during the entire session. However, the therapist only intervened by helping the children pressing the right button when prompted by Artur during the initial instructions. Since the children could not read, it would otherwise have been difficult to understand which button to press (the buttons had only text, no icons).

All three children were positive to ARTUR. Only one could mention anything in particular that was good and that was he liked to practice pronunciation of the word "säl" /sɛːl/ (seal). Another child described the session as "difficult, but fun". The main disadvantage was that it was difficult to imitate the pronunciation. Two of the children appreciated the animation of the speech organs; the third thought that it was "strange". All thought that imitating the animation worked well.

These children had the same problems interpreting the black line representing the hard palate in the picture (see Figure 1).

All tried the four buttons, see Figure 1, during the instructions. However, none of them used them during the training session. The reason for this was probably that they could not read, and the buttons had no iconic representation, but also that they were new to CBSTs in general and ARTUR in particular.

Only one child managed to compare ARTUR with other CBSTs, and placed ARTUR as number two. "Kakadua", a program for creating stories was placed as number one. The main reason was the funny sound effects in Kakadua.

The two children that compared ARTUR to practicing fricatives with their speech therapist, considered the speech therapist to be better (the speech therapist in question was present during this interview).

4.3.1 Conclusions

For this younger group of children, the benefits of ARTUR in its present state are more limited. An interface directed to this age group must have more game-like features.

The conclusion clearly illustrates that an important requirement for a successful CBST system is that the user group is well-defined and that the training is adapted to the user's age and speech or articulation disorders. A previous study [11] suggests that this should be done by providing a general framework for articulation training, which should be adapted to each child by the speech therapist.

4.4 Accompanying adults

All accompanying adults were fascinated of ARTUR, even though we explained that it was a Wizard-of-Oz-test, and we were only faking the system. A suggestion from one of the teachers was that the children often wanted to have a goal in their assignments, and a way of knowing how they were doing. In this case, just knowing the number of words and seeing a progress bar would be an improvement. Also a reward when the task is finished was appreciated.

4.5 Wizard impressions

The Wizards subjective impression of the training sessions can be summarized as:

The children did improve their pronunciation during the session by following the instructions from Artur.

The ten feedback options provided a too crude mapping of the pronunciation errors encountered, and it was sometimes impossible to catch smaller errors with the available feedback. As a fall-back solution, when such errors occurred, Artur was made to give only an encouragement ("Good try!", "It sounds better now" or "You're really good!") and the same word was repeated again.

The solution to this problem would however not be to introduce a finer feedback matrix, but rather to have a confidence score on the determined feedback (regardless of if the decision is made by a human Wizard as here, or automatically by the system), as the Wizard sometimes felt that the feedback instructions were too detailed. Instead of giving precise information on how to correct the articulation, a lower confidence score should generate feedback at a higher and looser level, e.g., "Almost, but think about how you place your tongue tip".

The feedback given should depend on the previous performance on the current and preceding words. In the current implementation, the same articulatory feedback instruction was given each time a specific error occurred. This must be changed in order to get both an enhanced training of the current word and a more rewarding variation between training words. If the student repeats the same error on the same training word, the system must go to a second level of feedback, where either more or less focus is placed on the error made, depending on how crucial the error is. Repeating the exact same feedback would quickly bore the student. In the current study, the Wizard tackled the problem again by giving an encouragement rather than feedback on a repeated error. In addition, a limit was set to avoid repeating the same word more than three times. Repeated errors between different training words should be handled similarly; if it is an important feature, additional focus should be placed on the feedback concerning this feature, and if it is less crucial, the system should tend to accept this particular deviation for the time being and focus on the most important. Conversely, if the child has only small difficulties with an articulation, finer details should be given in the feedback.

The focus of the training session must be clear both for the student and the future automatic feedback decision algorithm; if it is on one particular articulation or on the best production of the training words. The Wizard found for several subjects that they did not have difficulties with the initial fricative, but made other errors in the word. Due to the set-up of the training session, he was unable to give feedback on these other mispronunciations. Giving feedback on one specific part of a word may also result in a better production of this part, but a worse mispronunciation over the entire word, as other parts are altered. This is not an artifact, but a result of the focus of the training, that should first be on separate articulations in the word, and then later on the entire word, when its constituent parts are mastered.

The functionality of the interaction buttons may have been conceptually clear to the children, as they stated in the interviews, but in the practical use they did cause some confusion. One reason was that the implementation required the buttons to be disabled (which was signaled with grey shading) when Artur spoke. Some users tried to interrupt Artur during his utterances by pressing the button, and as nothing would happen, the user may conclude that the button was not working. A related problem was encountered for the "Show difference" button, which may only be pressed when the user had made an error (as there would otherwise not be any difference to show). A few users tried to press this button on other occasions and got no response. It must hence be self-evident to the user when available buttons may be used, or it must be possible to interrupt the program.

4.5.1 Conclusions

The classification matrix of pronunciation errors needs to be supplemented with a set of higher level, and less detailed, feedback instructions, when the articulation error falls between the defined categories.

The amount and detail of feedback should adapt on-line to the user's performance.

The focus of the training session should be stated explicitly and feedback should only be given on these articulations. However, other pronunciation errors should be logged in order to be able to suggest adequate training foci for subsequent sessions.

More varied encouragement, in particular such that are less related to the actual pronunciation task, are needed.

The usefulness of the interaction buttons was not evident. A supplement may be that the virtual tutor takes the initiative for additional feedback, if this is judged to be needed, e.g., "*Would you like to see the difference?*" or "*Would you like to see me say the word slowly?*".

5. DISCUSSION

The main goal with this study was to get early feedback from the potential users of CBST, not to measure any efficiency of the system. Although the number of people involved in this study is small, and no objective measures of longitudinal improvements have been made, there are certain observations that we believe to be of general interest.

First, it is clearly possible to make an interface to a speech training system that can be used by children on their own without training or instructions. This is a necessary first step if the finished system is to be used in the children's home.

Second, if such a system existed (for example a completely functional ARTUR), speech therapists and the older children would regard it as a major support in their training.

Third, any system that gives feedback based on classifications similar to the ones described in Table 1 needs a systematic handling of uncertainty, lack of finer granularity and possible misclassification.

Fourth, children like computer games, and this should be exploited in computer assisted learning.

5.1 Future work

The interviews in this study and in [11] have shown that there is a clear need for motivating factors in the program to inspire the children with enthusiasm for the training. We will hence carry out a study of motivational features in commercial pedagogical computer games and the most promising features will be tested in the system. Our goal is to create a training situation where the child is playing a game rather than focusing hard on the articulation. We believe that this will create a more stimulating training situation, in which the child is willing to spend more time, and thus getting more practice.

6. ACKNOWLEDGEMENTS

This study was carried out within the ARTUR project, funded by the Swedish Research Council. The Wizard of Oz study was sponsored by an award to the second author from the Christian Benoît Association. The Centre for Speech Technology is supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The authors would like to thank the speech therapists and children participating in the interviews. Fredrik Winberg for comments on earlier versions of this paper.

7. REFERENCES

[1] Ahlberg, J. Model-based Coding – Extraction, Coding and Evaluation of Face Model Parameters, PhD Thesis 2002, Linköping University, Sweden.

[2] Adams, F.R., Crepy, H., Jameson, D., and Thatcher, J. IBM products for persons with disabilities Global Telecommunications Conference, 1989, and Exhibition. 'Communications Technology for the 1990s and Beyond'. GLOBECOM '89, IEEE, 27-30 Nov. 1989, Vol. 2, 980 - 984

[3] Barker, J. & Berthommier, F. Evidence of correlation between acoustic and visual features of speech, Proc. of the Int. Congress of Phonetical Sciences 1999, pp. 199-202.

[4] Beskow, J. Talking Heads – Models and Applications for Multimodal Speech Synthesis, 2003, Ph.D. Thesis, KTH, Sweden. ISBN 91-7283-536-2.

[5] Bunnell, H.T. Yarrington, D.M. & Polikoff, J.B. STAR: articulation training for young children, In: Int. Conference on Spoken Language Processing 2000, Vol.4, pp. 85-88.

[6] De la Torre, F. & Black, M.J. Robust parameterized component analysis: applications to 2D facial modeling, Proceedings of the sixth European Conference on Computer Vision 2002, pp 653-669.

[7] Engwall, O. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model, Speech Comm., 2003, Vol. 41 (2-3), pp. 303-329.

[8] Engwall, O. Introducing visual cues in acoustic-to-articulatory inversion (submitted).

[9] Engwall, O. Wik, P., Beskow, J., and Granström B. Design strategies for a virtual language tutor, In: Int. Conference on Spoken Language Processing 2004, Vol. III, pp. 1693-1696.

[10] Erber N.P. Visual perception of speech by deaf children: recent developments and continuing needs, Journal of Speech and Hearing Disorders, 1974, Vol. 39:2, pp. 178-185.

[11] Eriksson E., Bälter O., Engwall O, Öster A-M and Kjellström H. Design Recommendations for a Computer-Based Speech Training System Based on End-User Interviews. To be published in proceedings of SPECOM 2005.

[12] IPA, International Phonetic Alphabet. URL: http://www2.arts.gla.ac.uk/IPA/index.html Last retrieved July 28 2005.

[13] Kroos, C., Kuratate, T. & Vatikiotis-Bateson, E., Listen to the face - measuring the face kinematics of speech from video sequences. Proceedings of the 5th Int. Seminar on Speech Production, pp. 341-344.

[14] Markides, A. Lipreading: Theory and practice, Journal of Brittish Association of Teachers of the Deaf, 1989, Vol. 13:2, pp. 29-47.

[15] Massaro, D.W. and Light, J. Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss, Journal of Speech, Language and Hearing Research, Vol. 47, April 2004, pp. 304-320

[16] Neti C, Potamianos G, Luettin J, Matthews I, Glotin H, Vergyri D, Sison J, Mashari A, and Zhou J. Audio-visual speech recognition, Final Report from Workshop 2000 Audio-Visual Speech Recognition.

[17] OLP, (2003) OLP Home. URL: http://www.xanthi.ilsp.gr/olp/default.htm Last retrieved: Nov. 24 2004

[18] Öster, A-M. (1996) "Clinical applications of computer-based speech training for children with hearing-impairment". Proceedings of ICSLP-96, 4th Int. Conference on Spoken

Language Processing, Philadelphia, USA, Oct 1996; pp. 157-160.

[19] Öster, A-M. House D., Green P., Testing a new method for training fricatives using visual maps in the Ortho-Logo-Pedia project (OLP), Phonum 9, Fonetik 2003, Umeå, pp. 89-92.

[20] Reeves B. and Nass C. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. University of Chicago Press. ISBN 157586053.

[21] Rubin, J. Handbook of Usability Testing, John Wiley & Sons, Inc, 1994, ISBN 0-471-59403-2

[22] Soleymani, A.J.A., McCutcheon, M.J. & Southwood, M.H. Design of speech mentor (SIM) for teaching speech to the hearing impaired. In: Proceedings of the 1997 Sixteenth Southern Biomedical Engineering Conference, pp. 425-428

[23] Vicsi K., Roach P., Öster A.-M., Kacic Z., Barczikay & Tantoa A., Csatári F. & Bakcsi Zs., Sfakianaki A. A multilingual teaching and training system for children with speech disorders, Int. Journal of Speech technology, 2000, Vol. 3, 289-300.

[24] Watson, C., Reed, D., Kewley-Port, D. and Maki D., The Indiana Speech Training Aid (ISTRA). Comparisons Between Human And Computer-Based Evaluation of Speech Quality, Journal of Speech and Hearing Research, June 1989, Vol. 32, pp. 245-251

[25] WHO. http://www.who.int/classifications/icd/en/

[26] Wiepert, S.L., Mercer, V.S. Effects of an increased number of practice trials on Peabody Developmental Gross Motor Scale scores in children of preschool age with typical development. Pediatric Physical Therapy 2002, Vol. 14, pp. 22-28.