Taylor & Francis
Taylor & Francis Group

Designing the user interface of the computer-based speech training system ARTUR based on early user tests

OLOV ENGWALL*†, OLLE BÄLTER‡, ANNE-MARIE ÖSTER† and HEDVIG KJELLSTRÖM‡

†Centre for Speech Technology (CTT), School of Computer Science and Communication,
KTH (Royal Institute of Technology), Sweden
‡Interaction and Presentation Laboratory, School of Computer Science and Communication,
KTH (Royal Institute of Technology), Sweden

This study has been performed in order to evaluate a prototype for the human – computer interface of a computer-based speech training aid named ARTUR. The main feature of the aid is that it can give suggestions on how to improve articulations. Two user groups were involved: three children aged 9 – 14 with extensive experience of speech training with therapists and computers, and three children aged 6, with little or no prior experience of computer-based speech training. 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, but that more motivational factors should be introduced. The granularity of the mesh that classifies mispronunciations was satisfactory, but the flexibility and level of detail of the feedback should be developed further.

Keywords: Computer-based speech training system; User interface; Wizard of Oz test; Participatory design

1. Introduction

Imitation and self-correction are important factors in speech and language learning, and a hearing impairment or a mother tongue from a different language group may hence lead to communication problems caused by difficulties in hearing what is wrong with one's own pronunciation. This applies both to children with hearing or language disabilities and adult second language learners.

Children who are born with a severe auditory deficit have a limited acoustic speech target with which to imitate and compare their own articulation, and other senses must replace the auditory feedback that hearing children use when they learn to speak. Severely and prelingually hearing-impaired children have to rely on the limited visibility of phonetic features in learning oral speech and on orosensory motor control in maintaining speech movements. These children seldom develop speech spontaneously, but their speech can be developed through a structured training with speech therapists, who use the visibility of speech articulation, reading, tactile sensations and, if possible, residual hearing (Dodd 1974, Ling 1976, Levitt and Geffner 1987, Oller 2000).

Second language (L2) learners face a similar challenge when distinctions in the L2 do not exist in the mother tongue (L1). The pronunciation of a word varies greatly depending on the speaker (age, dialect, gender, mood, health, etc.), the situation (formal or informal, reading or talking, monologue or dialogue) and the context in which the word is pronounced. A very important process in children's language learning is thus to establish categorical perception, in which speech sounds are clustered into phonemes. This means that whereas inter-phonemic differences (i.e. acoustic differences between different phonetic sounds, e.g. 'r' and 'l') lead to classification into different categories, intra-phonemic differences (i.e. differences in the same phonetic sound between different contexts or speakers) are accepted as variability within the category.

*Corresponding author. Email: engwall@kth.se

Behaviour & Information Technology
ISSN 0144-929X print/ISSN 1362-3001 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/01449290600636702

Categorical perception is a prerequisite for spoken communication, but it also implies that distinctive contrasts in the L2 that are non-distinctive in the L1 can cause problems. A major challenge in pronunciation training in a new language is therefore to make the student aware of these unfamiliar distinctions.

1.1 *Computer-based speech therapy and computer-assisted pronunciation training*

Computer-based speech therapy (CBST) systems is one of several important tools for speech therapists when practising with hearing- or language-impaired children. Examples of such systems are SpeechViewer (Adams *et al.* 1989), Box of Tricks (Vicsi *et al.* 2000), Indiana Speech Training Aid (ISTRA) (Watson *et al.* 1989), Speech Illumina Mentor (SIM) (Soleymani *et al.* 1997), Speech Training, Assessment, and Remediation system (STAR) (Bunnell *et al.* 2000), and the OLP method (OLP 2003). Of these, both SpeechViewer and Box of Tricks are used extensively and acknowledged by speech therapists.

CBST has been shown to be very efficient, especially in the instruction phase of speech training (Öster 1996, Vicsi *et al.* 2000) since a computer-assisted aid is capable of offering a child immediate and meaningful visual feedback of various distinctive contrasts. By this technique it might also be easier for the therapist to instruct and explain what is wrong and what is correct in the child's speech (Osberger *et al.* 1981, Öster 1992).

Motor learning theory in speech development further indicates that accurate feedback and repeated practice are essential to establish automaticity and to transfer skills to untrained situations (Wiepert and Mercer 2002). This is the most important element in a speech therapy program but the most difficult for a therapist to carry out due to time constraints. The target production must be repeated and practised in a variety of contexts. To use CBST in this situation might be particularly helpful in motivating the student to undertake significant amounts of additional training (Öster *et al.* 2003), especially if it can be done without continuous supervision from the therapist (Eriksson *et al.* 2005).

Computer Assisted Pronunciation Training (CAPT), on the other hand, most often refers to pronunciation training in Computer Assisted Language Learning (CALL) of a second language. The potential market for CAPT systems is enormous (estimates of the global language learning market range from over USD 15 billion to USD 50 billion), but the major breakthrough for CALL has yet to come, as existing products are hampered both by technological limitations and pedagogical issues in the feedback given (Neri *et al.* 2002a). We will mainly focus on speech training for children with hearing impairments or language disorders in the remainder of this article, but all major considerations about the human–computer interface that apply to CBST systems are also applicable to CAPT systems.

CBST and CAPT systems in general provide feedback by offering a success score or, more or (often) less, advanced visualization of the *acoustic* signal (waveforms, spectrograms or pitch curves). Neri *et al.* (2002b) made a thorough survey of existing CAPT systems and concluded that neither the score nor the visualization of the acoustic signal was sufficient for adequate unsupervised pronunciation training. This conclusion is all the more true for a hearing-impaired child with limited notion of the acoustic targets. Hence it is often more fruitful to focus on visual or tactile properties of the pronunciation.

Massaro and Light (2003, 2004) used the talking head model Baldi (Cohen and Massaro 1993) as a virtual articulation teacher for Japanese students of English and for American hearing-impaired children, respectively. Baldi gave audiovisual instructions on how to produce difficult English sounds correctly and both groups were quite enthusiastic about the articulatory animations. The studies showed that the hearing-impaired children using the virtual tutor did indeed benefit from the audiovisual instructions, whereas the L2 students did not benefit to the same extent. Baldi does not, however, relate the instructions to the student's own pronunciation, nor is any feedback given. As not only imitation but also self-correction is an important factor 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 the target.

2. ARTUR—the ARticulation TUtoR project

Thus, in order to give students advice on how to improve their pronunciation a new CBST system, the ARticulation TUtoR (ARTUR) (Engwall *et al.* 2004) 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 spelling 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 advantage of ARTUR is that the feedback is given in the form of clear instructions on how to improve the articulation and also through animations of salient parts of these instructions. For example, if a user practising the r–l distinction pronounces '*Harry Potter*' as '*Hally Pottel*', Artur would reply, for example: '*That sounded more like Hally 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*', and he would show this difference in tongue tip positioning graphically.

The use of a talking head with internal parts is a key feature, whereby phonetic features that are hidden in a human speaker can be displayed. The perception of speech through normal lip-reading is difficult because many articulatory and acoustic features of speech are not easily visible. 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 a non-visible articulation (Erber 1974, Markides 1989). With a talking head, on the other hand, parts of the anatomy may be removed to display the manner and place of articulation (see figure 1 for an example).

One 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 (e.g. Massaro 1987, Summerfield 1987), learning will be more robust and efficient with multimodal training than with either modality alone (as shown by Massaro and Light 2004, for example).

The development of the ARTUR system involves several different research areas (figure 2). Many of the components are still at a very early stage of development as they require efforts at the research frontier in each domain.

Detection of mispronounced speech. The input to the system is the user's utterances, and the goal is to detect erroneous deviations between the target and the user's pronunciation, based mainly on acoustic data. This is not a trivial task, as the system must allow variability on the one hand (the goal is for the student to achieve a good pronunciation, not exactly the same as in the target), but on the other hand it must have a definition of a correct pronunciation in order to be able to detect those that differ from what a human listener would accept.

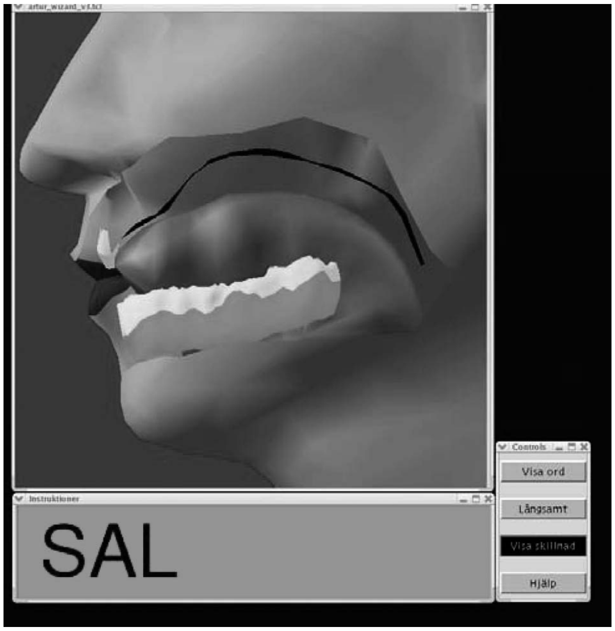

Figure 1. The user interface giving articulatory feedback. Top left: side view of the talking head model, with a part of the chin removed to make the intra-oral articulation visible. Bottom left: training word. Right: user control buttons.

The solution to this problem is a combination of a theoretical framework and statistical methods in speech recognition. The theoretical framework consists of preconceived notions about the errors that the user is prone to make based on the language background and the type of hearing impairment of the user, which would be known to the system either because the user has been logged when using the system previously or because the information has been provided by the user, a therapist or a teacher. Such information will assist the system in setting up the classification categories for the speech recognition (e.g. that hearing-impaired children will often confound the pronunciation of 's' and 'sh' if they have a hearing-impairment affecting higher frequencies). The statistical methods consist in training the speech recognizer with both correct (normal hearing or native speakers, depending on whether the system is to be used for CBST or CAPT) and deviant (hearing-impaired or L2 speakers) pronunciations, as done previously by Deroo *et al.* (2000).

As the expected input from the user is generally known in CBST and CAPT (the exercises typically consist of repeating words or sentences, practising a specific articulation or answering closed questions) it can be compared to a target utterance using forced alignment (i.e. the speaker's utterance is matched sound by sound with the target text to obtain the best possible fit between the audio and the text). Previous work in this field includes, for example, the reading coach by Mostow *et al.* (1994) and the automatic pronunciation error detection by Jo *et al.* (1998).

Extracting visual speech information from a video of the face. In speech recognition tasks, such as the mispronunciation detection described above, visual and acoustic information are complementary. The strong influence of the visual information in human speech understanding is demonstrated by McGurk and MacDonald (1976), and Neti *et al.* (2000) showed that visual data improves the performance of automatic speech recognition as well, especially in noisy conditions. Furthermore, there are important correlations between jaw and lip configuration and speech acoustics (Barker and Berthommer 1999), which means that video images can be used to increase the robustness of the mispronunciation detection. As an example, a very common error for L2 learners of Swedish or French is to replace a rounded vowel (e.g. /y/, as in the Swedish 'by' = 'village' or French 'lu' = 'read') by the

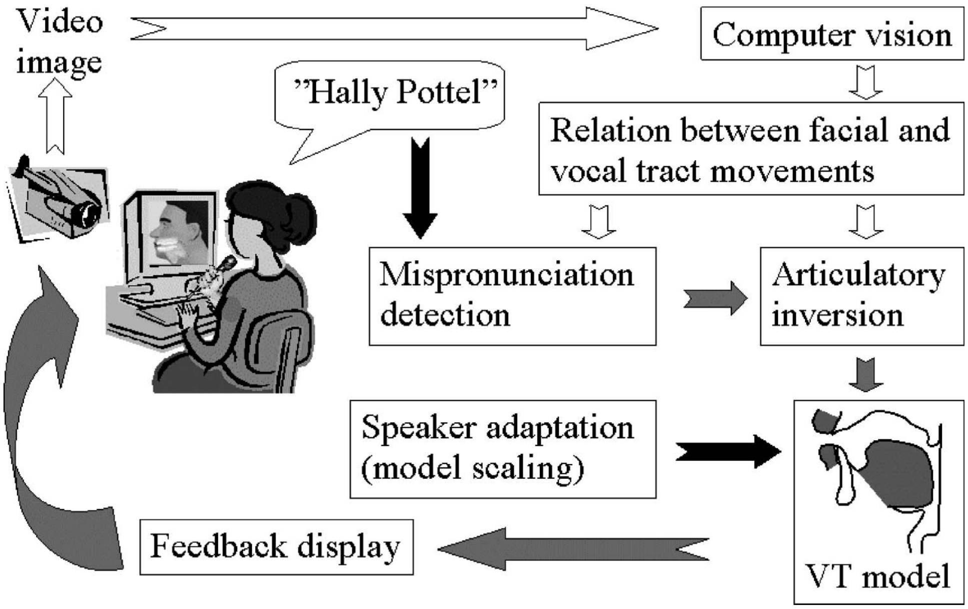
Video image
Computer vision
"Hally Pottel"
Relation between facial and vocal tract movements
Mispronunciation detection
Articulatory inversion
Speaker adaptation (model scaling)
Feedback display
VT model

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

counterpart with spread lips (i.e. /i/ as in 'bi' = 'bee' or 'lit' = 'bed', respectively), if the lip rounding contrast does not exist in the mother tongue. This error may be difficult to detect automatically on the acoustic level, but it is straightforward to spot in the visual data.

Neti *et al.* (2000) divided the methods for extraction of lip and jaw parameters into three main groups: high-level geometric methods that track the lip contours, low-level pixel-based methods, and hybrids of these. The high-level methods represent the face in terms of the shape of the inner and outer mouth contours. Parameters could then be the mouth opening, the jaw opening, the distance between the corners of the lips, and the lip rounding. These methods are susceptible to changes in appearance due to lighting changes and change of speakers. However, they only make use of a small fraction of the information present in the image, and will deliver no information at all if the contour tracking fails. The low-level methods instead represent the face in terms of the pixel values in the video image. For example, a set of 'basis images'—images with typical mouth deformations—could be learned from a set of training images of different mouth shapes. Any new face image can be represented as a linear combination of those basis images. The parameters describing the mouth shape are then the weights of each basis image in the linear combination. In ARTUR, we are currently working along this direction.

Articulatory inversion. The next step is to recreate the user's motion of the face and vocal tract from the speech signal and facial parameters. Acoustic to articulatory inversion (or speech inversion) is one of the major remaining challenges in speech technology research. The main difficulty is that there is no unique mapping between the acoustic and articulatory domains, and a large number of vocal tract shapes may produce the same speech sound. The problem of finding the articulatory representation from the speech signal is hence under-determined, as there are more unknowns that need to be determined than there are input data available. It is therefore necessary to introduce constraints that are both sufficiently restrictive and phonetically realistic, in order to eliminate false solutions. These constraints are traditionally derived from speech production models of the vocal tract, but no inversion system has yet been able to reliably find unique solutions to a realistic acoustic input. Existing inversion techniques are mainly applicable to vowels and sequences of vowels of one speaker (Laprie and Ouni 2002, Ouni and Laprie 2003) and substantial efforts are hence necessary to achieve a general-purpose inversion to be useful for more speakers and more varied utterances. One approach that is exploited in ARTUR is to employ video images to extract information about visible articulators, i.e. to perform visio-acoustic to articulatory inversion. As stated above, there is a significant correlation between the face and the tongue positions, and facial data can hence improve the articulatory inversion substantially (Engwall 2005).

Articulatory model. The user's correct articulations are synthesized using the models of the face (Beskow 2003) and vocal tract (Engwall 2003) developed at KTH (figure 3). The role of talking head models is to represent the human

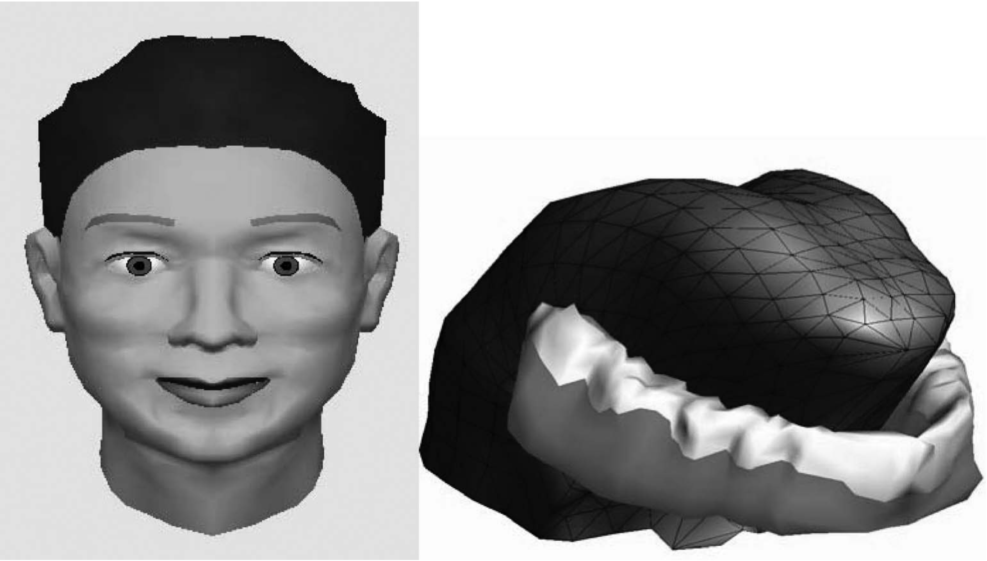

Figure 3. (a) The talking face model ARTUR. (b) The tongue and jaw model. The wireframe mesh shows the underlying structure.

speech production as closely as possible, while making the required simplifications to make the model computationally useable. The KTH models are based on concepts first introduced by Parke (1982), defining a set of parameters that deform a static 3D-wireframe mesh by applying weighted transformations to its vertices. The parameters for the face are *jaw opening, jaw shift, jaw thrust, lip rounding, upper lip raise, lower lip depression, upper lip retraction* and *lower lip retraction*.

The 3D vocal tract model (VT model) consists of three-dimensional mesh structures of the tongue, jaw, palate and vocal tract walls. The tongue model is controlled by articulatory parameters changing the *jaw height, tongue dorsum raise, body raise, tip raise, tip advance* and *width*. These parameters were defined through a statistical component analysis of Magnetic Resonance Imaging (MRI) data of one Swedish reference subject producing 13 Swedish vowels and 10 consonants in three symmetric vowel-consonant-vowel contexts (Engwall 2003). As the acquisition time of 43 s required the subject to artificially sustain the articulations, Electromagnetic Articulography (EMA), Electro-palatography (EPG) and real-time MRI measurements of the same subject (Engwall 2003, 2004a) were used to adjust the articulations to those occurring in normal speech and to obtain information on articulatory kinematics.

Adaptation of the model to the user. As the shape of the face and vocal tract varies between individuals, the articulatory inversion requires that the model be adapted automatically to each new user. Engwall (2004b) showed that four articulatory measures in a midsagittal MR Image of a new subject were enough to adapt the whole 3D tongue model to that user, with an accuracy of 1.5 mm in the midsagittal plane and 1.7 mm for the 3D tongue. The adaptation was trained using a statistical analysis of MRI data from 9 subjects, and the aim was to be able to do rescale the model based on acoustic input and initial information on the speaker's age and gender.

Feedback display. The output, an articulatory representation of the training utterance, is crucial for the success of the system (Neri *et al*. 2002b). It is important that the feedback is comprehensible, useful and motivating for the student. It is a delicate task to present just enough information about the system without overwhelming the user and at the same time give enough information so that the user can understand the difference between his/her performance and the goal. The development of the interface therefore requires expertise in several areas including human–computer interaction, speech therapy, pedagogy, and computer science. A natural method for the development of the feedback display is to use participatory design that includes experts in all areas as well as the students and teachers (Muller *et al*. 1997). A previous study (Eriksson *et al*. 2005) used structured interviews with speech therapists and their students to summarize their perceived needs, requirements and wishes for CBST systems. The current Wizard of Oz study was carried out as a subsequent 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 would influence the requirements on the components. The study further served the purpose of collecting audio and video data that would be important training material for the mispronunciation detection.

3. Method

3.1 *The Wizard of Oz set up*

The set up of the Wizard of Oz system differed from the planned future system in the aspects shown in figure 4. The mispronunciation detection and the articulatory inversion were performed by a phonetically trained human Wizard (the first author); some system tasks were disabled and the audio and video recordings were stored to create an audio-visual database, to be used for system training later on. The user interface, shown in figure 1, consisted of one window displaying the virtual tutor Artur (implemented as a virtual face of an approximately 10 year-old boy, see figure 3) and his articulatory feedback animations, one text window showing the training words and sub-titling of all Artur's utterances (as an additional support for hearing-impaired users) and one set of interaction buttons.

The test was carried out in a studio, where the student was placed alone in a sound-proofed room in front of the computer screen with the ARTUR interface, while the Wizard controlled the training session from an adjacent room (see figure 5). A window between the two rooms allowed the Wizard to observe the training session.

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. A microphone was fitted on the collar of the subject's sweater, and the audio signal was transferred to the Wizard's headphones and recorded on disk. During the training session, the inputs from the user were vocal (uttering the training words) or with the mouse to click the interface buttons, whereas the Wizard controlled the feedback and encouragements using a cordless keyboard with shortcut keys.

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 (Beskow 2003, Engwall 2003). During the introduction, the student was given the opportunity to test the interaction buttons shown in figure 1: 'Show word' (Visa ord), 'Slow' (Långsamt) and 'Show difference' (Visa skillnad). Pressing 'Show word' resulted in a repetition of the training word articulation animation; '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 of 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 /s/ or /ɧ/ (voiceless velar fricative with rounded lips. Refer to IPA, 1999 for definitions of the phonetic symbols) 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), 'sida' vs. 'skida' (page vs. ski), 'sol' vs. 'kjol' (sun vs.

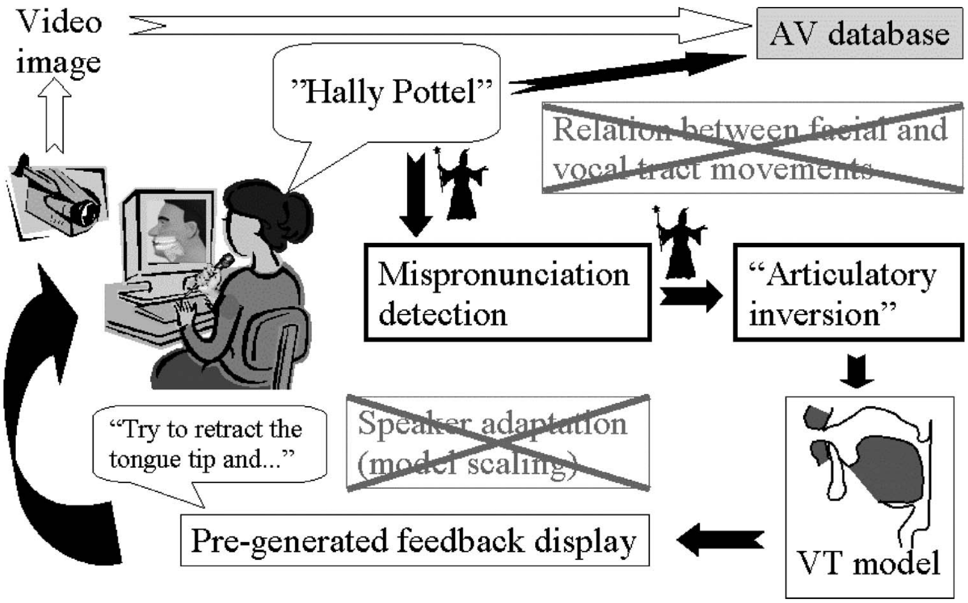
Video image
AV database
"Hally Pottel"
~~Relation between facial and vocal tract movements~~
Mispronunciation detection
"Articulatory inversion"
"Try to retract the tongue tip and..."
~~Speaker adaptation (model scaling)~~
Pre-generated feedback display
VT model

Figure 4. Schematic overview of the Wizard of Oz version of the ARTUR system. The wizard symbol indicates tasks where the wizard replaced the automatic system.

skirt), ‘sula’ vs. ‘skjul’ (sole vs. shed), ‘sylt’ vs. ‘skylt’ (jam vs. sign), ‘så’ vs. ‘sjå’ (sow vs. tough job) ‘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. Minimal pairs of these two fricatives were chosen since they are often confounded by hearing-impaired children, due to the high frequency of the frication noise and the difficulty of seeing the place of articulation from looking at the speaker’s face.

After each student utterance, the Wizard selected the appropriate feedback from the 10 different feedback options in table 1, three encouragement utterances (‘*Good try!*’, ‘*It sounds better now*’ or ‘*You’re really good!*’) 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 student’s tongue, which could be judged to be incorrect in positioning or manner of articulation, or both. Note that the generated feedback depended on the training word, as, for example, the detection of a word-initial/ɕ/ (alveolo-palatal fricative) should result in feedback indicating that the articulation should be more forward if the training word began with /s/, but more retracted if it started with /ɧ/.

Feedback, when the student made an error, 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, for example, for the training word ‘sal’: ‘*Almost! Now you said “tal”. Try to lower the tongue tip, so that the air can pass.*’

3.2 *Interviews*

After the test, the student and the interface researcher (the second author) went to another room separate from the test laboratory for an interview about the interface to ARTUR. The separate interview room was to avoid effects from having the tested system present during the interview (Reeves and Nass 1996). The interviews were semi-structured (Rubin 1994) using an interview script with open-ended questions, but with the possibility of probing the interviewee further if needed. This is especially suitable for interviews with children, where the interviewer has the opportunity 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.

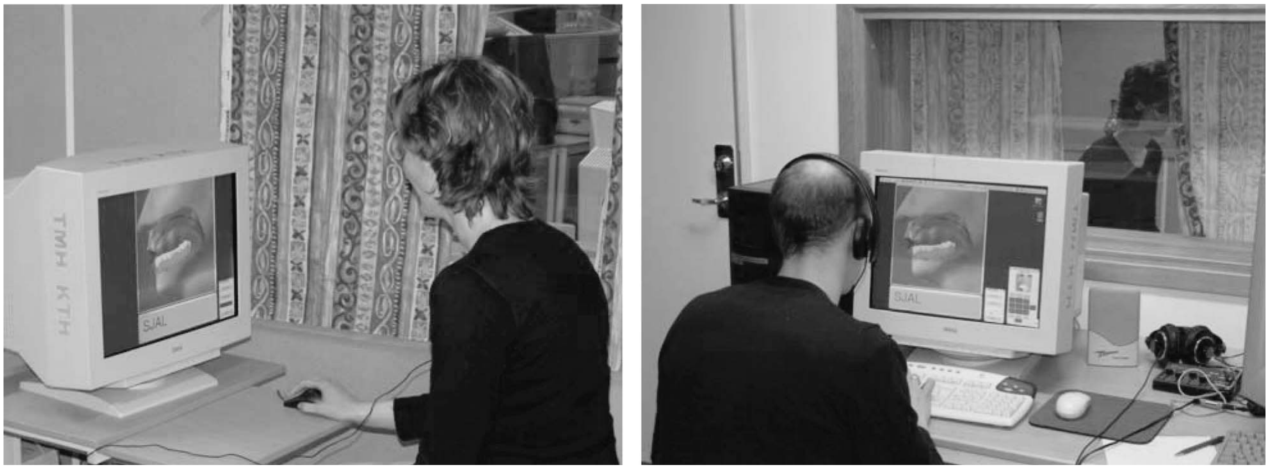

Figure 5. The Wizard of Oz set-up of ARTUR. Left: the user with the articulatory feedback interface, mouse and microphone. Right: Wizard, with the same user interface, keyboard with shortcut keys and the observation window looking into the test studio.

Table 1. The feedback options available to the Wizard, 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 /ħ/; too backward.
	Lisp	No audible fricative.	Fricative made between tongue edges and the teeth.	

A test session lasted approximately 10 min and the following interviews another 15 min.

3.3 *Test subjects*

Two user groups were interviewed: three children aged 9–14 with extensive experience of speech training and CBST systems, and three children aged 6 at 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's mother tongue is Persian, but is fluent in English and Swedish, and has experience of second language learning as well as CAPT.

None of the three older children (9–14) had any hearing difficulties but all had language disorders. According to the International Classification of Diseases (ICD 10), which is available online at http://www.who.int/classifications/icd/en/ (accessed 22 July 2005), 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 varying degrees. 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 one adult accompanied the child and assisted when the answers or questions were unclear. The accompanying adult was a parent for the first child, and one speech therapist and one teacher for the other two children, respectively.

The group of three younger children (6 years old) all had several years of experience of speech training but little experience of CBST. 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 of making descriptions. These children had their speech therapist sitting next to them during the test. This was mainly to support the children during Artur's introduction. 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 explain to the child that the purpose of the test was to scrutinize the system, 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, with the meetings taking place in the child's home or school.
3. To make the interviewer a familiar person for the child.
4. To make the interviewer (and hence the Wizard) aware of the child's strengths and weaknesses before the test and interview.

For the group of 6 year olds, the interview script was adapted in order to better fit the age group. This was accomplished by replacing some words with simpler versions (e.g. 'do the same' instead of 'imitate'), and by making it possible to express opinions by pointing at iconic faces (see figure 6). For comparisons of ARTUR to other CBST systems, paper slips were prepared with text and iconic pictures representing the different systems. These slips were given to the child to sort them in order of preference.

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, but it was concluded that the reason that these children had not seen a see-through picture of the articulation was not that it would be unnatural at this stage of training, but simply that no such picture was available. The speech therapists said that if they had had a picture 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 interviews for the different users are presented below.

4.1 *Adult second language learner*

The adult testing the system as a pre-study 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 practise a pronunciation a few times before being evaluated by the system*'.

A CBST or CAPT system used by a normal hearing person should not assume that feedback is necessary after

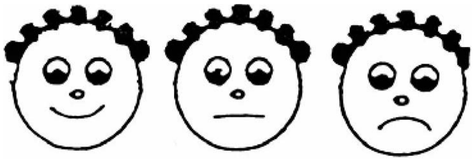

Figure 6. Iconic faces used for expressing opinions in the interviews with the younger children.

each attempt. In learning a language with speech sounds deviating greatly from the mother tongue, the students may hear that their own pronunciation is wrong, and may want several attempts to get it right. In such cases, it should be possible to skip the feedback by pressing a button, for example. Allowing the student to judge him- or herself when the pronunciation is good enough to be scrutinized by the virtual tutor could be an important part of the process to make the student aware of his/her own articulation and deviations from the target utterance.

4.2 *Children aged 9–14*

All three children were very positive about ARTUR. They described it as '*very intelligent and good and so*'. They concluded that the best part was the oral and written instruction on how to improve the pronunciation by moving the tongue more forward or backward. Their main objection against the current version of the system was the limited number of sounds that could be practised, as they had already established quite a stable /s/–/ɧ/ contrast during previous speech training.

The animation of the speech organs varied in popularity. One of the children said that it was '*really good*'; one complained of some technical flaws and the third was not too 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 it is where the air is coming*?'

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

All understood the function of the four buttons 'See again', 'Slow', 'Show difference' and 'Help' (see figure 1). However, the buttons were not often used during the training session. When asked, they explained that 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 CBST systems, all preferred ARTUR. One child said that it was '*twice as good as SpeechViewer and Box-of-Tricks*'. The main reason for this was the correction instructions given on the pronunciation. The articulatory feedback was hence appreciated by the children, but it should be acknowledged that the possibility of giving adequate, detailed instructions relies heavily on the performance of the speech recognizer in an automatic system. If the recognition fails, this may result in erroneous instructions, which would be a severe drawback of the system. Strategies to avoid such errors are hence essential, as discussed further in section 4.5.

The children indicated that there were, however, features that were better in other CBST systems, such as the possibility of practising more varying sounds and scoring for correct pronunciation.

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

4.3 *Children aged 6*

The younger children had their speech therapists present during the entire session, but the therapist only intervened by helping the children press 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, see figure 1).

All three children were positive about ARTUR. Only one could, however, mention anything in particular that was good and that was that he liked to practise 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 as the older children in interpreting the black line representing the hard palate (see figure 1).

All tried the four interaction buttons shown in figure 1 during the instructions, but 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 the children were new to CBST in general and ARTUR in particular.

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

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

4.4 *Accompanying adults*

All accompanying adults were fascinated by ARTUR, even though we explained that it was a Wizard of Oz test and we were only simulating parts of 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. A reward when the task is finished would also be appreciated.

4.5 *Wizard impressions from the training sessions*

The Wizard's subjective impression of the training sessions was that the children did improve their pronunciation during the session by following the instructions from Artur, but the short training session did not generate sufficient data for any quantitative evaluation of student performance and progress. The objective of this study was not, however, to estimate the objective impact of the speech training with ARTUR, but to interview the users about their impression of the system, in particular the human–computer interface. Subsequent studies with a larger training corpus or more repeated practice should test the system's ability to improve students' pronunciation.

The 10 feedback options in table 1 provided too crude a mapping of the pronunciation errors encountered, and it was sometimes impossible to catch smaller errors with the available feedback. As a fallback solution, when such errors occurred, Artur was made to say one of the three encouragements and the same training word was repeated again, without giving any articulatory feedback. The solution to this problem would not be to introduce a finer feedback matrix, but rather to have a confidence score on the determined feedback (regardless of whether 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 and did not exactly correspond to the error that the student was judged to have made. Instead of giving precise information on how to correct the articulation, a lower confidence score should generate feedback at a higher and looser level, for example '*Almost, but think about how you place your tongue tip*', to avoid giving erroneous detailed feedback.

The feedback given should further depend on the previous performance on the current and preceding words. In the tested implementation, the same feedback instruction was given each time a specific error occurred. This must be changed in order to achieve 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 should switch 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 judged to be. Repeating the 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 it, 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 feature. Conversely, if the child only has a little difficulty with an articulation, finer details should be given in the feedback.

Furthermore, the focus of the training session must be clear both for the student and the automatic feedback decision algorithm, i.e. if it is on one particular articulation or on the best pronunciation of the entire training word. 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 other mispronunciations. Giving feedback on the fricative part of a word did sometimes also result in a better articulation of this part, but a worse mispronunciation over the entire word, as other parts were altered. Note that this is not an artefact of the Wizard of Oz set up, but a result of the focus of the training, which should first be on separate articulations in the word, and only later on the entire word, when its constituent parts have been mastered.

The functionality of the user buttons may have been conceptually clear to the children, as they stated in the interviews, but in practical use they did cause some confusion. One reason was that the implementation required the buttons to be disabled (which was signalled with grey-shading) when Artur spoke. Some users tried to interrupt Artur's utterances by pressing a button, and as nothing would happen, the user may have concluded that the button was not working. A related problem was encountered for the 'Show difference' button, which may only be used when the user has 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 received no response. It must hence be evident to the user when available buttons may be used, or it must be possible to interrupt the program.

4.6 *Conclusions from the interviews*

For the older children, the idea of giving feedback on how to alter the articulation seems fruitful, especially the written or oral instructions. For the group of younger children, the benefits of ARTUR in its present state are more limited. This 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 (Eriksson *et al.* 2005) suggests that this should be done by providing a general framework for articulation training, which should be adapted to each child by the responsible speech therapist.

When it comes to the animations of the articulatory movements the results are mixed, but a learning effect may make the animations more useful for the students as they become more familiar with it. However, we believe that the animation speed of the articulatory feedback needs to be altered to separate the articulation that is being practised from the rest of the word. The animation now shows a slow but natural pronunciation 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 more slowly and exaggerated while the remaining parts are shown at normal speed.

The usefulness of the interaction buttons was not evident. The lack of use of the functions activated by buttons may have been caused by the novelty of these functions, but may also be explained by the fact that these children had only minor problems with the pronunciation of the training words. A supplement may be that the virtual tutor takes the initiative for additional feedback if this is judged to be needed, for example, '*Would you like to see the difference?*' or '*Would you like to see me say the word slowly?*'

The representation of the hard palate clearly needs improvement.

More game-like features would increase the interest from the children to practise with the system, especially concerning the younger children. A wider variety of encouragement, in particular related less to the actual pronunciation task than, for example, how many training words are left, is also needed.

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

The amount and detail of feedback should adapt online 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.

5. Discussion

The main goal of this study was to obtain early feedback from the potential users of CBST. Although the number of subjects in the study is small, and no objective measures of longitudinal improvements were made, there are certain observations that we believe are of general interest.

First, it is clearly possible to make an interface to a CBST system that can be used by children on their own without prior 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 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 misclassifications.

Fourth, keeping the user motivated is a key factor for successful longer training sessions, and a more sophisticated encouragement protocol should be implemented. As children like computer games, we strongly believe that speech training with game-like features would be beneficial.

6. Future work

The interviews in this study and in Eriksson *et al.* (2005) 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, concentrating on visual and multimodal features that can be used for hearing-impaired children. 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 achieving an articulation. This will create a more stimulating training situation, in which the child is willing to spend more time and thus has more practice. As an example, we would like to replace the 'repeat after me' training of two phonemes described above with an assault course game, where the player is riding a bicycle moving upwards on the screen and has to avoid obstacles by saying a word that turns the bicycle left (e.g. 'sal') or right (e.g. 'sjal'). The pair of words will be changed when a sufficient number of correct pronunciations have been achieved. Similar games do exist in SpeechViewer version 3, but the crucial difference is that the game in ARTUR would be centered on articulatory feedback. The correct front articulation will always be displayed in the upper left corner of the screen, and the correct back articulation in the upper right, for reference (as Artur always shows articulations facing left, as in figure 1, left will be associated with a more frontal articulation and right with more backward). At each turn, ARTUR has to classify the pronounced word into five categories: one or the other correctly produced (one category each), neither (one category), or a failed attempt at one of the two training words (one category each). In the last two cases, the game would be momentarily paused and articulatory feedback given. The category 'neither' is essential as a means of avoiding problems such as happy shouts interrupting the game. We also envisage implementing games for more than two training words, e.g. Tetris- (three articulations) or PacMan-like (four articulations) games.

The display of the palate has already been re-implemented as a consequence of the user comments. The main difficulty is that the palate is essential as a reference for tongue–palate contact and distance, but must at the same time not occlude the tongue. The new representation instead displays the tongue moving in a black oral cavity with the palate and cheek as the delimiter of this cavity.

We see continuous user testing and feedback as a cornerstone in the development process of ARTUR and will hence make repeated tests and user studies and interviews of either Wizard of Oz tests or fully automatic versions of the system as the work progresses.

Acknowledgements

The ARTUR project is funded by the Swedish Research Council, and this Wizard of Oz study was sponsored by the third Christian Benoît award given to the first author by 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 study, and Fredrik Winberg and David House for comments on earlier versions of the manuscript.

References

Adams, F.R., Crepy, H., Jameson, D. and Thatcher, J., 1989, *IBM Products for Persons with Disabilities. Global Telecommunications Conference and Exhibition. 'Communications Technology for the 1990s and Beyond', GLOBECOM '89*, IEEE, **2,** pp. 980–984.

Barker, J. and Berthommier, F., 1999, Evidence of correlation between acoustic and visual features of speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*, 1–8 August 1999, San Francisco, CA, pp. 199–202 (Berkeley: University of California).

Beskow, J., 2003, Talking heads—models and applications for multimodal speech synthesis. Ph.D. Thesis, KTH, Sweden.

Bunnell, H.T., Yarrington, D.M. and Polikoff, J.B., 2000, STAR: articulation training for young children. In *Proceedings of the 6th International Conference on Spoken Language Processing*, **4,** 16–20 October 2000, Beijing, China, pp. 85–88 (Beijing: China Military Friendship Publish).

Cohen, M.M. and Massaro, D.W., 1993, Modelling coarticulation in synthetic visual speech. In *Speechreading by Humans and Machines: Animation*, N. Magneat-Thalmann and M. E. Hennecke (Eds), pp. 139–156 (Tokyo: Springer Verlag).

Deroo, O., Ris, C., Gielen, S. and Vanparys, J., 2000, Automatic detection of mispronounced phonemes for language learning tools. In *Proceedings of the 6th International Conference on Spoken Language Processing*, **1,** 16–20 October 2000, Beijing, China, pp. 681–684 (Beijing: China Military Friendship Publish).

Dodd, B., 1974, The acquisition of phonological skills in normal, severely subnormal and deaf children. Doctoral dissertation, University of London.

Engwall, O., 2003, Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication*, **41,** pp. 303–329.

Engwall, O., 2004a, From real-time MRI to 3D tongue movements. In *Proceedings of the 8th International Conference on Spoken Language Processing*, 4–8 October 2005, Jeju Island, Korea, pp. 1109–1112 (Korea: Sunjin Printing Co.).

Engwall, O., 2004b, Speaker adaptation of a three-dimensional tongue model. In *Proceedings of the 8th international Conference on Spoken Language Processing*, 4–8 October 2005, Jeju Island, Korea, pp. 465–468 (Korea: Sunjin Printing Co.).

Engwall, O., 2005, Introducing visual cues in acoustic-to-articulatory inversion. In *Proceedings of Interspeech*, 5–9 September 2005, Lisbon, Portugal, pp. 3205–3208 (Rundle Mall, Australia: Casual Productions Pty Ltd.).

Engwall, O., Wik, P., Beskow, J. and Granström, B., 2004, Design strategies for a virtual language tutor. In *Proceedings of the 8th International Conference on Spoken Language Processing*, 4–8 October 2005, Jeju Island, Korea, pp. 1693–1696 (Korea: Sunjin Printing Co.).

Erber, N.P., 1974, Visual perception of speech by deaf children: recent developments and continuing needs. *Journal of Speech and Hearing Disorders*, **39,** pp. 178–185.

Eriksson, E., Bälter, O., Engwall, O., Öster, A-M. and Kjellström, H., 2005, Design recommendations for a computer-based speech training system based on end-user interviews. In *Proceedings of the 10th International Conference on Speech and Computer*, 17–19 October 2005, Patras, Greece, pp. 483–486 (Moscow: Moscow State Linguistics University).

IPA, 1999, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet* (Cambridge: Cambridge University Press).

Jo, C-H., Kawahara, T., Doshita, S. and Dantsujiy, M., 1998, Automatic pronunciation error detection and guidance for foreign language learning. In *Proceedings of the 5th International Conference on Spoken Language Processing*, 30 November–5 December 1998, Sydney, Australia, pp. 289–292 (Rundle Mall, Australia: Casual Productions Pty Ltd.).

Laprie, Y. and Ouni, S., 2002, Introduction of constraints in an acoustic-to-articulatory inversion method based on a hypercubic articulation table. In *Proceedings of the 7th International Conference on Spoken Language Processing*, 16–20 September 2002, Denver, CO, pp. 211–214 (Rundle Mall, Australia: Casual Productions Pty Ltd.).

Levitt, H. and Geffner, D., 1987, Communication skills of young hearing-impaired children. *ASHA Monographs*, **26,** pp. 123–158.

Ling, D., 1976, *Speech and the Hearing-impaired Child: Theory and Practice* (Washington, D.C.: The A.G. Bell Association for the Deaf, Inc.).

Markides, A., 1989, Lipreading: theory and practice. *Journal of British Association of Teachers of the Deaf*, **13,** pp. 29–47.

Massaro, D., 1987, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Hillsdale: Lawrence Earlbaum Associates).

Massaro, D.W. and Light, J., 2003, Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, 1–4 September 2003, Geneva, Switzerland, pp. 2249–2252 (Rundle Mall, Australia: Casual Productions Pty Ltd.).

Massaro, D.W. and Light, J., 2004, Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language and Hearing Research*, **47,** pp. 304–320.

McGurk, H. and MacDonald, J., 1976, Hearing lips and seeing voices. *Nature*, **264,** pp. 746–748.

Mostow, J., Roth, S., Hauptmann, A. and Kane, M., 1994, A prototype reading coach that listens. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pp. 785–792 (Seattle, WA: American Association for Artificial Intelligence).

Muller, M.J., Haslwanter, J.H. and Dayton, T., 1997, Participatory practices in the software lifecycle. In *Handbook of Human–computer Interaction*, M.D. Helander, T.K. Landauer and P.V. Prabhu (Eds), pp. 255–297 (Amsterdam: Elsevier Science).

Neri, A., Cucchiarini, C. and Strik, H., 2002a, Feedback in computer assisted pronunciation training: when technology meets pedagogy. In *Proceedings of CALL Conference CALL professionals and the future of CALL research*, 18–20 August 2002, Antwerp, Belgium, pp. 179–188 (Antwerp: University of Antwerp Press).

Neri, A., Cucchiarini, C., Strik, H. and Boves, L., 2002b, The pedagogy–technology interface in Computer Assisted Pronunciation Training. *Computer Assisted Language Learning*, **15,** pp. 441–467.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A. and Zhou, J., 2000, Audio-visual speech recognition. Final report from Workshop 2000 Audio-Visual Speech Recognition, 12 October, Baltimore, MD.

Oller, K., 2000, *The Emergence of the Speech Capacity* (Mahwah: Lawrence Erlbaum).

OLP, 2003, Available online at: http://www.xanthi.ilsp.gr/olp/default.htm (accessed 24 November 2004).

Ouni, S. and Laprie, Y., 2003, A study of the French vowels through the main constriction of the vocal tract using an acoustic-to-articulatory inversion method. In *Proceedings of the 15th International Congress of Phonetic Sciences*, 3–9 August 2003, Barcelona, Spain, pp. 2901–2904 (Rundle Mall, Australia: Casual Productions Pty Ltd.).

Osberger, M.J., Moeller, M.P. and Kroese, J.M., 1981, Computer-assisted speech training for the hearing impaired. *Journal of the Academy of Rehabilitative Audiologists*, **14,** pp. 145–158.

Öster, A-M., 1992, The speech of deaf children—phonological assessment as a basis for speech-training. Thesis work for the Licentiate Philosophy degree in Phonetics, University of Stockholm, 22 April 1992.

Öster, A.-M., 1996, Clinical applications of computer-based speech training for children with hearing-impairment. In *Proceedings of the 4th International Conference on Spoken Language Processing*, 3–6 October 1996, Philadelphia, PA, pp. 157–160 (New Castle, Delaware: Citations Delaware).

Öster, A-M., House, D., Hatziz, A. and Green, P., 2003, Testing a new method for training fricatives using visual maps in the Ortho-Logo-Pedia project (OLP). In *Proceedings of Fonetik 2003, Phonum 9*, Reports in Phonetics, Umeå University, Sweden, pp. 89–92.

Parke, F.I., 1982, Parameterized models for facial animation. *IEEE Computer Graphics*, **2,** pp. 61–68.

Reeves, B. and Nass, C., 1996, *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places* (Chicago: University of Chicago Press).

Rubin, J., 1994, *Handbook of Usability Testing* (New York: John Wiley & Sons Inc).

Soleymani, A.J.A., McCutcheon, M.J. and Southwood, M.H., 1997, Design of speech mentor (SIM) for teaching speech to the hearing impaired. In *Proceedings of the Sixteenth Southern Biomedical Engineering Conference*, 4–6 April 1997, Biloxi, MS, pp. 425–428 (Piscataway, NJ: IEEE Engineering in Medicine and Biology Society).

Summerfield, Q., 1987, Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by Eye: The Psychology of Lip-reading*, B. Dodd and R. Campbell (Eds), pp. 3–52 (Hillsdale: Lawrence Erlbaum Associates).

Vicsi, K., Roach, P., Öster, A-M., Kacic, Z., Barczikay, Tantoa, A., Csatári, F., Bakcsi, Z. and Sfakianaki, A., 2000, A multilingual teaching and training system for children with speech disorders. *International Journal of Speech technology*, **3,** pp. 289–300.

Watson, C., Reed, D., Kewley-Port, D. and Maki, D., 1989, The Indiana speech training aid (ISTRA). Comparisons between human and computer-based evaluation of speech quality. *Journal of Speech and Hearing Research*, **32,** pp. 245–251.

Wiepert, S.L. and Mercer, V.S., 2002, 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*, **14,** pp. 22–28.