Intelligent Animated Agents for Interactive Language Training

Ron Cole, Tim Carmell, Pam Connors, Mike Macon, Johan Wouters,
Jacques de Villiers, Alice Tarachow
Center for Spoken Language Understanding, Oregon Graduate Institute, Portland, Oregon

Dominic Massaro, Michael Cohen, Jonas Beskow†
Perceptual Science Laboratory, University of California, Santa Cruz

Jie Yang, Uwe Meier, Alex Waibel
Interactive Systems Laboratory, Carnegie Mellon University, Pittsburgh, Pennsylvania

Pat Stone, George Fortier, Alice Davis, Chris Soland
Tucker-Maxon Oral School, Portland Oregon

Abstract

This report describes a three-year project, now eight months old, to develop interactive learning tools for language training with profoundly deaf children. The tools combine four key technologies: speech recognition, developed at the Oregon Graduate Institute; speech synthesis, developed at the University of Edinburgh and modified at OGI; facial animation, developed at University of California, Santa Cruz; and face tracking and speech reading, developed at Carnegie Mellon University. These technologies are being combined to create an intelligent conversational agent; a three-dimensional face that produces and understands auditory and visual speech. The agent has been incorporated into the CSLU Toolkit, a software environment for developing and researching spoken language systems. We describe our experiences in bringing interactive learning tools to classrooms at the Tucker-Maxon Oral School in Portland, Oregon, and the technological advances that are required for this project to succeed.

1. Introduction

In September 1997, the Oregon Graduate Institute in collaboration with the University of California, Santa Cruz (UCSC), and Carnegie Mellon University (CMU), was awarded an NSF Challenge grant for a three year project to develop interactive tools, technologies and applications for learning and language training with profoundly deaf children.

The goal of the project is to provide teachers, students, parents and other interested individuals with state of the art tools and technologies for interactive learning and language training. These tools and technologies are being integrated into the CSLU Toolkit, a software environment for building and researching interactive media systems [1].

To date, language-training applications have been developed using an animated conversational agent, called Baldi [2]. The agent is represented by an animated 3D face that produces visual speech—facial movements of the lips, jaw, cheeks, and tongue during speech production [1]. The visual speech is aligned to a speech waveform to produce an animated talking face.

The linguistic input to the facial animation program is a string of phonemes, their durations, and a pitch contour (used to control eyebrow movements). So far, the linguistic information has been provided by the Festival TTS system [3]. Recently, we added the ability to synchronize recorded natural speech with the animated face. The speaker records an utterance and types in the words. The system then produces a time-aligned phonetic transcription that is used to drive the synthetic face.

When facial animation and speech generation are combined with computer speech recognition, students are able to have limited conversations with Baldi. These conversations are created by CSLU's toolkit developers and by educators at the Tucker-Maxon Oral School, using the toolkit's graphical authoring tools.

2. Language Training at Tucker-Maxon Oral School (TMOS)

Language training software is being developed and tested in collaboration with the educators and deaf students at the Tucker-Maxon Oral School [4] in Portland, Oregon. The students in this school are profoundly deaf. Hearing aids, cochlear implants, or a combination of both are used to enhance their hearing.

During the first six months of the project, our main challenge has been to adapt and extend the toolkit to the

† Jonas Beskow is visiting UC Santa Cruz from KTH.

Intelligent animated agents for interactive language training. In STiLL: ESCA Workshop on Speech Technology in Language Learning (pp. 163-166). Stockholm, Sweden, May, 1998.

needs of the teachers and students. To this end, we have followed principles of participatory design, in which the users of the software participate in all phases of its design and development.

In the summer of 1997, four educators from TMOS (the executive director, speech pathologist, and two teachers) took a short course at OGI on building spoken dialogue systems using the CSLU Toolkit. During the course, each of the educators learned to build interactive spoken language systems, and developed a language training application for their course project.

2.1 Classroom Experiences

In the fall of 1997, we installed four Pentium II PCs, donated by Intel, at TMOS. Two of the computers were installed in a classroom of 11-13 year-olds, and one in a classroom of 8-10 year-olds. The fourth was placed in the speech lab for use by the speech therapist with children from both classes as well as some younger children. These educators have integrated the toolkit into their instruction in unique ways. Tim Carmell, Pam Connors and Alice Tarachow have collaborated with the educators to develop specific applications.

George Fortier's Class

In the class of the oldest students, instructor George Fortier has taken full advantage of the toolkit's authoring environment to create interactive multi-media systems. These dialogues introduce and/or review concepts and vocabulary from the social studies and science curricula.

For example, he created a geography lesson by scanning in a drawing that depicted a mountain range, lake, river, stream, plateau, waterfall, coast, and ocean. When a student starts this program, one of the geographic features is selected randomly and highlighted. The conversational agent, Baldi, asks: "What landform is this?" A correct answer by the student causes the system to display another highlighted landform and ask another question. If the child does not know the answer she says, "Help." Baldi then says the word three times at a slightly slower rate (90% of normal) and asks the question again.

If the system does not recognize the student's speech Baldi asks her to try again. After two recognition failures, Baldi again says the word three times. This "correct speech routine" continues until the student's speech is recognized. Because of the children's varying levels of speech ability the teacher sets the recognition threshold differently for each child. This threshold is increased daily so that the student's speech quality must improve for recognition to occur.

This program has been an effective learning tool. None of the six students knew all of these landforms before using the application. After four days of working individually, all 6 students knew the names of each landform and could say them with such clarity that the system accepted their answer on the first or second trial.

Mr. Fortier has created 15 different applications for his class. Working on a home PC in the evening, he uses the toolkit's authoring tools to create sophisticated interactive dialogues that incorporate different graphics at specific points in the dialogue. All of these applications exercise and test language comprehension, speech production and content learning skills.

Alice Davis' Class

Alice Davis, the instructor of the younger students, has created a variety of applications for different content areas to supplement her curriculum. For example, she has built modules for the study of math, spelling, reading, listening comprehension and writing. Most recently, in preparation for the school spring program, she has used the toolkit as a way for the children to listen to and practice reciting their own haiku.

To test math concepts, Ms. Davis created interactive math quizzes. For example, the child was shown a picture of seven bears in a forest. Baldi would then say: "There are seven bears in the forest. Three bears ran away. How many bears are left?" When the child produced the correct answer, she received the next question. Applications of this type sharpen listening skills and speech production skills, while testing knowledge of content.

In another tutorial, Ms. Davis created an application that combined speech, pictures and text to present a story written by the children. These young authors were then able to listen to Baldi tell their story accompanied by Ms. Davis' drawings displayed at appropriate times during the story. At the conclusion of the session, the children answered comprehension questions composed by the teacher. During this activity, children were able to view a dynamic multi-media presentation of their story, and then interact with the system to demonstrate comprehension and production skills.

Speech Therapy

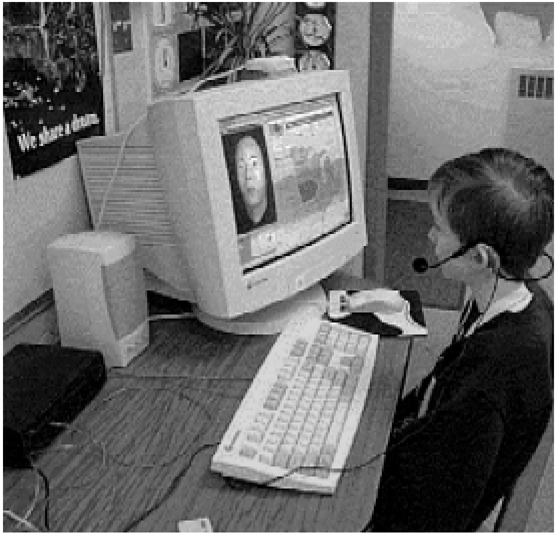
The speech therapist, Chris Soland, in collaboration with Tim Carmell, developed a series of listening drills based on minimal pair distinctions. For example, the screen shows a picture of "mail" and "veil". The system then says one of the words, and the student uses a mouse to select the picture that corresponds to the word that she perceived.

After completing the identification task the child begins a same-different discrimination task. In this segment, the system produces a pair of words, e.g., "veil, veil" or "mail, veil," and the student says "same" or "different" after pair. In each application, the system produces immediate visual feedback and informs the child of her score at the end of the lesson.

Once the child has finished the listening activities, she is prompted by the application to record her best production of the target words. The child has the opportunity to listen to her production, record it until satisfied, and then save it to a file for the speech therapist to review.

The children enjoy these challenging tasks, and often ask to repeat the lesson after receiving a low score. One nine-year-old girl asked to repeat one lesson three times until she had a perfect score of 25 correct. She then asked the teacher to print out her score so she could show it to the principal.

Figure 1: A student at Tucker-Maxon Oral School using



the CSLU toolkit for learning and language training.

2.2 Summary of Initial Experiences

The introduction of the CSLU toolkit and the animated agent has produced immediate benefits to the teachers and students at the Tucker-Maxon Oral School. The educators report that the new learning tools have expanded their ability to teach vocabulary, practice and refine speech skills, introduce and reinforce academic subject matter, and refine and enhance listening skills.

The initial acceptance of the learning tools greatly exceeded the expectations of the toolkit developers. In the initial months of the project, the Windows 95/NT port of the toolkit was not stable, and the software caused frequent system crashes. Moreover, some of the technologies that are most important for language training are not yet available, such as the ability to recognize children's speech accurately, or the ability to display the articulators and their movements during speech production. The initial success of the project has been due in large part to the patience, enthusiasm and creativity of the TMOS teachers, and the graduate students who helped them adapt the tools to their needs.

3. Creating New Technologies for Language Training

In this section, we describe the work that is now underway in each of the four areas of auditory visual speech recognition and generation. The technological advances described below will be integrated into the CSLU Toolkit during the summer recess, and will be ready for use in language training applications at TMOS during the fall semester, 1998.

3.1 Speech Recognition

One of the main limitations of the learning tools is the poor recognition accuracy for kids' speech. Because of this, applications have been limited to a few words or phrases, and even then, egregious errors occur; i.e., the child produces an excellent utterance, and the system gives it a low recognition score, and responds "Sorry, please try again." The system performs poorly on kids' speech because the recognizer was trained on adult speech.

To overcome this limitation, we are collecting a kids' speech corpus at the Forest Grove school district in Oregon. We are collecting a set of phonetically balanced words and phrases, and samples of conversational speech, from about 200 students at each grade level, from kindergarten through 10th grade. When this data collection is complete, we will transcribe the utterances, and train recognizers that will work well with children's speech.

3.2 Speech Reading

The goal of the speech reading software being developed at CMU is to capture and interpret the student's facial movements during speech production [5]. Visual speech information will be combined with the acoustic signal to produce more accurate speech recognition, and to provide more accurate feedback to the student on the results of their speech production.

The face and lip tracking system has been has been ported to Windows 95/NT for integration into the CSLU Toolkit. The first application of the system will be to provide visual timing information to the students.

A major problem for deaf children is learning the temporal characteristics of speech. Providing this timing information visually could be a valuable tool. To provide this information to language learners, the system will locate and track the child's face and interpret the visual speech. Baldi will then reproduce the child's visual speech using the child's voice. A juxtaposed image of Baldi will then produce the utterance correctly. The child will be able to play these sequences at different speeds to observe the differences and practice the correct pronunciation.

3.3 Facial Animation

The talking head developed at the University of California at Santa Cruz provides realistic visible speech that is almost but not yet as accurate as a natural talker. This approximation is suitable in most situations, but is being improved for language training. One of our current goals is aimed primarily at enhancing the quality of the synthetic visible speech.

It is essential that language training should benefit from the ability to show a variety of faces (e.g. faces of different ages, genders, and races). Given that the present synthesis algorithm is closely tied to a single facial model, we plan to generalize the visual speech algorithm to control additional models of the face.

We believe that sophisticated language use goes well beyond simply segmental speech, and thus we plan to synthesize paralinguistic information, such as emotion, eye movements, and head movements. To achieve realistic synthesis, it is necessary to obtain measurements of facial, lip, and tongue movements during speech production. These data will be used to develop more accurate speech articulation for our synthesis system.

For the perceiver's eye to instruct language perception and production, it is important to reveal speech articulators that go beyond what perceivers can see in a natural face. Because the face can be made transparent, a view of the inside of the mouth can be highly informative in language training. Thus it is important to have a veridical synthesis of the tongue, the hard and soft palates, and the teeth and gum area. We are using ultrasound, MRI, and electropalatography data and analyses carried out by Maureen Stone at Johns Hopkins and University of Maryland, to guide the creation and control of the tongue.

An important problem that we are solving is that of collision detection, so that the appropriate contact between the tongue and the hard palate and teeth can be simulated. It is also necessary to change the shape of the tongue appropriately upon contact since it is an organ that maintains constant volume. We have obtained some very high-resolution models of internal speech articulation structures (e.g. teeth, gums, palate), which we are processing to reduce the number of polygons for real-time rendering. The soft palate will be modeled as being either open or closed for nasal and nonnasal segments. Morphing between these two endpoints appears to be a promising technique for dynamic synthesis.

3.4 Speech Generation

Our work on text-to-speech synthesis aims to provide more flexibility and control to the teachers for designing language-training applications.

Using a system for voice conversion based on short samples of user training data [6], the identity of the synthetic voice can be modified. In the future, we will be able to produce child voices using this technique. Our eventual goal is to have any user produce a short passage of speech, and have the system then generate speech from text in the users' voice.

In articulation training applications, it would be desirable to introduce hyperarticulation of words or sounds being practiced. To provide this capability to application developers, we will develop and implement a "What You See Is What You Hear" (WYSIWYH) editor that allows teachers to describe in intuitive, high-level terms how the prompts are to be rendered. This will be an integrated part of the authoring tools' dialog "prompt" box.

As a mechanism for specifying the information to the synthesizer, we are using the emerging standard markup language from the SABLE consortium, which originated from the markup language STML [7]. The editor window will save the highlighted, annotated text using this format. Using this marked up text also provides a mechanism for specifying facial and vocal expressions of emotion, hyperarticulation, and any other stylistic renderings of speech.

Acknowledgements

This work was supported in part by NSF CHALLENGE grant CDA-9726363, ONR/DARPA grant N00014-94-1-1154, NSF CARE grant EIA-9996075, Public Health Service grant PHS R01 DC00236), National Science Foundation grant 23818, and Intel Corporation. The CSLU Toolkit was developed with support from NSF, ONR and DARPA. We thank Irv Nikolai and the Forest Grove teachers and students for their help in collecting kids' speech.

References

[1] Information about the CSLU toolkit can be found at http://www.cse.ogi.edu/CSLU.

[2] Massaro, D. W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. Cambridge, MA: MIT Press. (Information about the facial animation software, developed at the University of California, Santa Cruz, is available at http://mambo.ucsc.edu/psl/pslfan.html)

[3] Black, A. and Taylor, P. (1997). Festival Speech Synthesis System: System documentation (1.1.1) Human Communication Research Centre Technical Report HCRC/TR-83. See http://www.cstr.ed.ac.uk for more information.

[4] Information about TMOS can be found at http://www.tmos.org/.

[5] Information about CMU's interactive systems lab can be found at: http://www.is.cs.cmu.edu/ISL.html.

[6] Kain, A. and Macon, M. W. "Spectral Voice Conversion for Text-to-Speech Synthesis," Proc. of International Conference on Acoustics, Speech, and Signal Processing, 1998.

[7] Sproat, R., Taylor, P., Tanenblatt, M. and Isard, A. "A markup language for text-to-speech synthesis," Proceedings of EUROSPEECH 97, Rhodes, Greece.