Massaro, D.W. (2003). A computer-animated tutor for spoken and written language learning. In Proceedings of the 5th International Conference on Multimodal Interfaces, (ICMI'03)(pp.172-175). Vancouver, British Columbia, Canada. New York: ACM Press.

A Computer-Animated Tutor for Spoken and Written Language Learning

Dominic W. Massaro
Department of Psychology
University of California, Santa Cruz
Santa Cruz, CA 95060 U.S.A.
1-831-459-2330
Massaro@fuzzy.ucsc.edu

ABSTRACT
Baldi, a computer-animated talking head is introduced. The quality of his visible speech has been repeatedly modified and evaluated to accurately simulate naturally talking humans. Baldi's visible speech can be appropriately aligned with either synthesized or natural auditory speech. Baldi has had great success in teaching vocabulary and grammar to children with language challenges and training speech distinctions to children with hearing loss and to adults learning a new language. We demonstrate these learning programs and also demonstrate several other potential application areas for Baldi®.

Categories and Subject Descriptors
J.4 [**Psychology**]

General Terms
Performance, Experimentation, Human Factors, Languages

Keywords: Language Learning, Facial and Speech Synthesis

1. INTRODUCTION
Speech as the primary communication medium has been a persistent goal to facilitate human machine interaction. Notwithstanding the challenges of naturally sounding speech synthesis, accurate speech recognition, and meaningful language understanding, some progress has been made. Even so, we cannot expect to achieve this goal in the near future. There are specific applications, however, in which humans can benefit from the speech and language produced and recognized by machine. We demonstrate the benefits of a computed-animated talking head for learning language, speech, and reading.

2. BALDI®[1]
The value of visible speech in face-to-face communication was the primary motivation for the development of Baldi, a 3-D computer-animated talking head [4]. The quality and intelligibility of his visible speech has been repeatedly modified and evaluated to accurately simulate naturally talking humans. Baldi's visible speech can be appropriately aligned with either synthesized or natural auditory speech. Baldi also has teeth, tongue, and palate to simulate the inside of the mouth, and his internal articulatory movements have been trained with electropalatography and ultrasound data from natural speech [3]. Combined with principles from linguistics, psychology and pedagogy, this technology has the potential to help individuals with language delays and deficits, as well as individuals learning a new language.

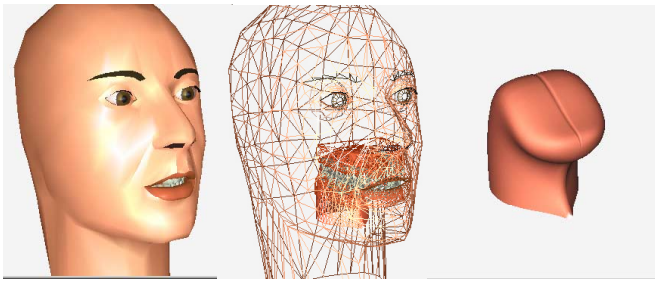

Figure 1. Baldi, a computer-animated talking head, in normal and wireframe presentations, and a close up of the tongue.

Computer-based instruction makes it possible to include embodied conversational agents rather than simply text or disembodied voices in lessons. There are several reasons why the use of auditory and visual information from a talking head is so successful, and why it holds so much promise for language tutoring [4]. These include a) the information in visible speech, b) the robustness of visual speech, c) the complementarity of auditory and visual speech, and d) the optimal integration of these two sources of information [4]. We will demonstrate several different applications utilizing Baldi to carry out language tutoring.

[1] Baldi is a trademark of Dominic W. Massaro

Table 1. Description of the 8 exercises available in the Language Tutor. Each exercise is optional and the specifications for each exercise can be made independently of the other exercises. For each exercise, the items can be randomly presented in a block of trials. The number of trial blocks is also specified independently for each lesson. The dialog and feedback for each exercise are also chosen by the coach creating the lesson. The feedback can include emoticons showing a happy or sad face.

Exercise	Description
Pre-Test	Baldi instructs the student to "click on the (word)" and the student is required to drag the computer mouse over the item that was just presented and click on it.
Presentation	One image is highlighted and Baldi tells the student "this is the (word)". Baldi then instructs the student to "show me the (word)" and the student is required to click on it. The student's response shows that they knew which image was being described.
Recognition	Baldi instructs the student to "click on the (word).
Readeing	The written text of each item is displayed in a separate area from the images. Baldi instructs the student to click on the written word corresponding to the highlighted image.
Spelling	One of the images is highlighted and Baldi asks the student to type the corresponding word.
Imitation	One of images is highlighted and Baldi names the item. The student is instructed to repeat the name Baldi had just said.
Elicitation	One of images is highlighted and Baldi asks the student to name it.
Post-Test	Baldi instructs the student to "click on the (word)".

3. LANGUAGE TUTOR

Baldi is the centerpiece in language tutoring software with Wizard and Player applications. Our Language Wizard is a user-friendly application that allows the composition of language,vocabulary and grammar lessons with minimal computer experience. Using the Language Wizard, the coach creates lessons that are individually tailored for each student. As described in Table 1, there are eight optional exercises that can be included. The Language Player engages the student in the lesson analogous to a video game [1,2]. Figure 2 shows a typical screen during a lesson. Images of the vocabulary items are presented on the screen next to Baldi as he speaks, as illustrated in Figure 2. Some of the exercises require the child to respond to Baldi's instructions such as "click on the cabbage", or "show me the yam", by clicking on the highlighted area or by moving the computer mouse over the appropriate image until an item is highlighted and then clicking on it. Two other exercises ask the child to recognize the written word and to type the word, respectively. The imitation and elicitation exercises ask the child to repeat after Baldi once he named the highlighted image and to name the highlighted image on their own. These utterances can be followed by the student hearing their production and/or Baldi saying the correct word(s).

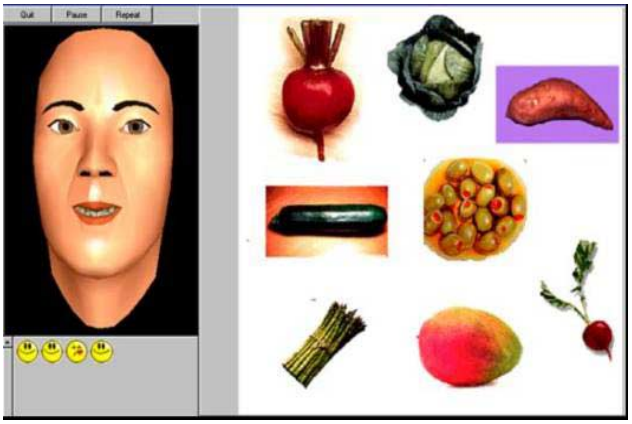

Figure 2. A computer screen from a vocabulary lesson, illustrating the format of the Language Player during one of the exercises. Each lesson can include Baldi, images of the vocabulary items, written text (not present in this exercise), and "stickers". In this application the students learn to identify fruits and vegetables. For example, Baldi says "Click on the beet". The student clicks on the appropriate region and feedback in the form of Baldi's spoken reaction and stickers (e.g., happy and disgusted faces) is given.

The existing program makes it possible for the students to 1) Observe the words being spoken by a realistic talking interlocutor (Baldi), 2) Experience the word as spoken as well as written, 3) See visual images of referents of the words, 4) Click on or point to the referent or its spelling, 5) Hear themselves say the word, followed by a correct pronunciation, 6) Spell the word by typing, and 7) Observe and respond to the word used in context.

Other benefits of our program include the ability to seamlessly meld spoken and written language, provide a semblance of a game-playing experience while actually learning, and to lead the child along a growth path that always bridges his or her current "zone of proximal development" [8]. The Wizard allows the coach to exploit this zone with individualized lessons, and with lessons that can bypass repetitive training when student responses indicate that material is mastered.

To evaluate the effectiveness of our Language Tutor, we carried out experiments based on a within student multiple baseline design where certain words were continuously being tested while other words were being tested and trained [2,5]. Although the students' instructors and speech therapists agreed not to teach or use these words during our investigation, it is still possible that the words could be learned outside of the Language Player environment. The single student multiple baseline design monitors this possibility by providing a continuous measure of the knowledge of words that are not being trained. Thus, any significant differences in performance on the trained words and untrained words can be attributed to the Language Player training program itself rather than some other factor.

Figure 3 gives the results of identification and production for one of eight students with hearing loss [7]. The results were highly

similar across the eight students. As can be seen in the figure, there was little knowledge of the test items without training, even though these items were repeatedly tested for many days. Once training began on a set of items, performance improved fairly quickly until asymptotic knowledge was obtained. This knowledge did not degrade after training on these words ended and training on other words took place. In addition, a reassessment test given about 4 weeks after completion of the experiment revealed that the students retained the items that were learned.

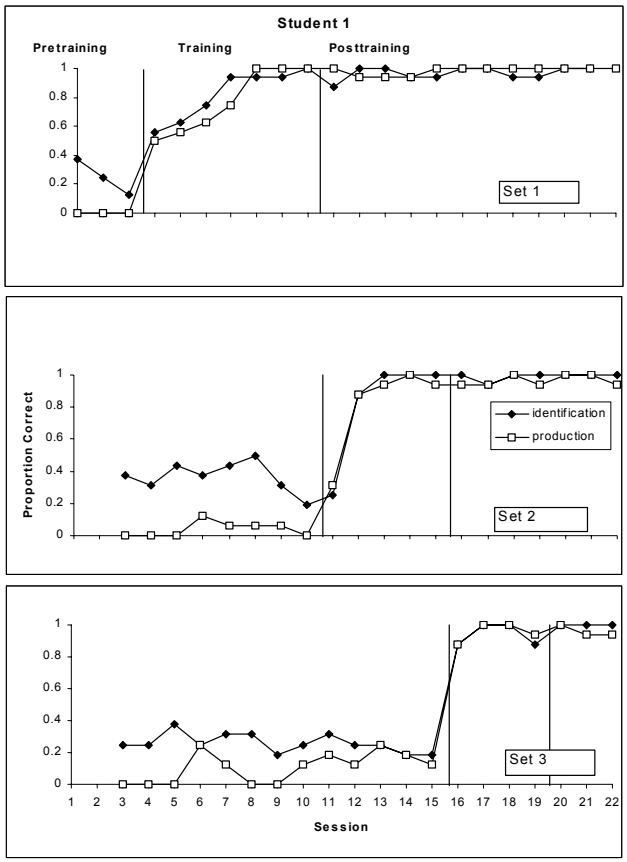

Figure 3. Proportion of correctly identified (solid black triangles) and correctly produced (empty white squares) items across the testing sessions for student 1. The training occurred between the two vertical bars. Once training began, identification performance increased dramatically, and remained accurate without further training.

Evaluation tests with both hard of hearing [1,5] and autistic [2] children indicated that they learned many new words, grammatical constructions and concepts, proving that the Language Player provided a valuable learning environment for these children. These experiments also showed that the program was responsible for the learning, and the learning generalized to new pictorial instances of the words outside of the learning situation.

4. SPEECH PRODUCTION TUTOR

Persons with hearing loss or those learning a nonnative language benefit from guided instruction in speech perception and production. Some of the distinctions in spoken language cannot be heard with degraded hearing--even when the hearing loss has been compensated by hearing aids or cochlear implants. Even if the distinctions are audible, these populations may have trouble perceiving and producing them. One reason is that many of the subtle distinctions among speech segments are not visible on the outside of the face. To illustrate these, the skin of our talking head can be made transparent so that the inside of the vocal tract is visible, or we can present a cutaway view of the head along the sagittal plane. The orientation of the face can also be changed to display different viewpoints while speaking. The auditory and visual speech can also be independently controlled and manipulated, permitting customized enhancements of the informative characteristics of speech. These features offer novel approaches to language training, permitting one to pedagogically illustrate appropriate articulations that are usually hidden by the face. Figure 4 gives some views used during several successful training experiments with both hard of hearing children [7] and adults learning nonnative speech contrasts in English [6].

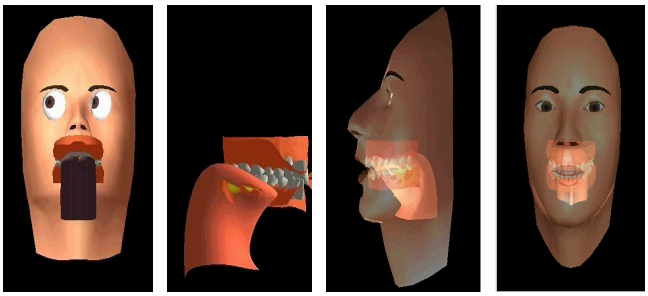

Figure 4. Four presentation conditions used in speech training (left to right: back view without a back of the head, sagittal view without the skin, side view, front view).

One of the original goals for the application of our technology was to use Baldi as a language and speech tutor for deaf and hard of hearing children. Baldi's technology seems ideally suited for improving the perception and production of English speech segments. Baldi can speak slowly, illustrate articulation by making the skin transparent to reveal the tongue, teeth, and palate, and show supplementary articulatory features such as vibration of the neck to show voicing and air expulsion to show frication. We [7] implemented these features in a set for language exercises, which involved viewing the segments being articulated in slowed speech under the views shown in Figure 4. Seven hard of hearing students between the ages of eight and thirteen were trained for six months on eight categories of segments (4 voiced vs. voiceless distinctions, 3 consonant cluster distinctions and 1 fricative vs. affricate distinction). Training included practice at the segment and the word level. Perception improved for each of the seven children and for each of the eight types of distinctions.

There was also significant improvement in production of these same segments. The students' productions of words containing these segments were recorded and presented to native English college students. These judges were asked to rate the intelligibility of a word against the target text, which was simultaneously

presented on the computer monitor. Intelligibility was rated on a scale from one to five (1:unintelligible, 2:ambiguous, 3:distinguishable, 4:unambiguous, 5:good/clear pronunciation). The children's speech production improved for each of the eight categories of segments. Speech production also generalized to new words not included in our training lessons. Finally, speech production deteriorated somewhat after six weeks without training, indicating that the training method rather than some other experience was responsible for the improvement that was found.

5. READING TUTOR

Psychological science has established a tight relationship between the mastery of written language and the child's ability to process spoken language. Many children with reading challenges also have deficits in spoken language perception. This difficulty with spoken language can be alleviated through improving children's perception of phonological distinctions and transitions, which in turn improves their ability to read and spell. Visible speech could only enhance the instruction of phonological awareness and therefore offers provide another dimension of information for the children to use in identifying segments and mastering phonological awareness.

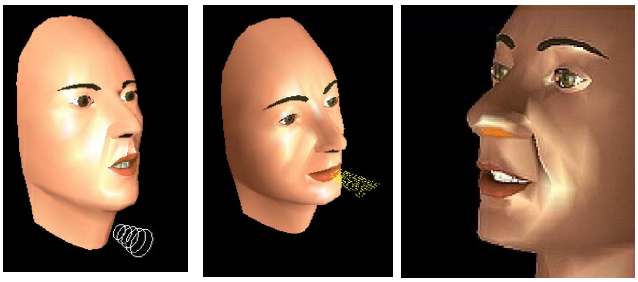

Figure 5. Supplementary features indicating from left to right, vocal cord vibration (rings from the neck), frication as in /s/ (dots from the mouth), and nasal as in /n/ (the red nostril color is not visible in the black and white illustration).

Baldi can be embellished to signal characteristics of the speech signal that could aid in the teaching of phonological awareness. Figure 5 illustrates some potential features that could be displayed along with the typical information given by visible speech. Baldi can vibrate his neck for voiced sounds, light up his nasal passage for nasal sounds, and show air rushing out of the mouth for fricative segments. Today, almost all personal computers have the capability to support a bimodal text-to-speech system, which would make it possible to incorporate bimodal speech in reading exercises. This treatment holds great promise in reading instruction, and we believe that adding visual speech will significantly enhance the positive results that have already been demonstrated with audible speech alone.

6. CONCLUSION

We have described several implementations of a language learning platform centered around Baldi, a computer-animated talking head. Baldi has served as a language tutor for deaf and autistic children in the learning of speech, vocabulary, and grammar. Baldi also has great potential for instruction in second language learning and in learning to read. We look forward to additional uses of Baldi in these areas and other domains such as the learning of social and conversational skills.

7. ACKNOWLEDGMENTS

The research and writing of the paper were supported by the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), a Cure Autism Now Foundation Innovative Technology Award, and the University of California, Santa Cruz. The technology and research resulted from the dedicated effort of many different members of the Perceptual Science Laboratory.

8. REFERENCES

[1] Barker, L. J. (in press). Computer-assisted vocabulary acquisition: The CSLU vocabulary tutor in oral-deaf education. *Journal of Deaf Studies and Deaf Education,* in press.

[2] Bosseler, A. and Massaro, D.W. (in press). Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning for Children with Autism. *Journal of Autism and Developmental Disorders,* in press. http://mambo.ucsc.edu/pdf/autism.pdf

[3] Cohen, M.M., Beskow, J., & Massaro, D.W. (1998). Recent developments in facial animation: An inside view. In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Eds.) *Proceedings of Auditory Visual Speech Perception '98*. (pp. 201-206). Terrigal-Sydney Australia, December, 1998. AVSP '98 (December 4-6, 1998, Sydney, Australia).

[4] Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* MIT Press: Cambridge, MA.

[5] Massaro, D.W., & Light, J. (in press). Improving the vocabulary of children with hearing loss, *Volta Review*, in press.

[6] Massaro, D.W., & Light, J. (2003). Read My Tongue Movements: Bimodal Learning To Perceive And Produce Non-Native Speech /r/ and /l/. *Eurospeech 2003-Switzerland (Interspeech). 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland.

[7] Massaro, D.W., & Light, J. (in press). Using Visible Speech for Training Perception and Production of Speech for Hard of Hearing Individuals. *Volta Review*, in press.

[8] Vygotsky, L. (1962). Thought and language. Cambridge, MA: MIT Press.